DETECTING OVERFITTING IN NEURAL NETWORKS DURING LONG-HORIZON GROKKING USING RANDOM MATRIX THEORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Training neural networks without overfitting is difficult; detecting that overfitting is difficult as well. We present a data-free Random Matrix Theory method for detecting the onset of overfitting from the trained weights alone. For each layer, we shuffle the weight matrix element-wise, $\mathbf{W} \to \mathbf{W}^{\mathrm{rand}}$, fit the shuffled matrix's empirical spectral distribution with a Marchenko–Pastur law, and identify large outliers beyond the finite-size right edge. We call these outliers *Correlation Traps*. In long-horizon grokking, traps form and grow in number and scale during *anti-grokking*, a post-generalization phase in which training accuracy remains high while test accuracy declines. Across an MNIST MLP, a modular-addition transformer, and a GPT2-style composition model, Correlation Traps distinguish anti-grokking from pre-grokking despite similar train and test behavior. Following Li and Sonthalia's spiked-regression account, a separated spike may be benign or harmful depending on its alignment with the learned function. We estimate this alignment without task data by passing random inputs through the trained model and measuring the Jensen–Shannon divergence between output logit distributions before and after suppressing each detected mode. The resulting scores recover the behavioral ranking obtained with in-distribution inputs, while continued suppression of the detected modes improves generalization more than an equal-count, equal-update-energy largest-weight edit. We also find Correlation Traps in foundation-scale language models, where their effect on generalization remains to be tested. These results identify randomized weight spectra as a data-free signature of anti-grokking and a practical screen for potential harmful overfitting.

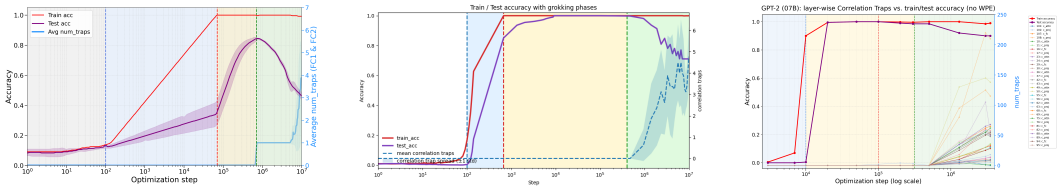
1 INTRODUCTION

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 (Power et al., 2022). Grokking has been studied across several architectures and tasks, including algorithmic tasks such as modular addition, computer vision models, and GPT-style transformers (Nanda et al., 2023; Wang et al., 2024). We extend this view to the long-horizon after grokking, where long-horizon 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 both regimes, training accuracy is high while held-out accuracy is poor. But they are structurally

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



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

Figure 1: **Grokking dynamics and Correlation Traps.** Reproduction of known grokking experiments, extended to long-horizon training. Train accuracy (red), test accuracy (purple), and Correlation Traps (blue) are shown for (a) an MNIST–MLP on MNIST, (b) Modular Addition (MODADD), and (c) GPT2–COMPOSE. Shaded regions denote pre-grokking (gray), grokking (yellow), and anti-grokking (green). Correlation Traps arise at the onset of late-stage test-accuracy decline and track it as it decreases.

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. In comparing different phases, this makes long-horizon grokking an excellent case for studying harmful overfitting.

To detect post-grokking overfitting structure, we introduce *Correlation Traps*. A Correlation Trap is a separated right-edge mode relative to an empirical MP/TW null fitted after entry-wise randomization of a layer weight matrix. For each layer, we shuffle the weights entry-wise, $\mathbf{W} \to \mathbf{W}^{\mathrm{rand}}$, fit the eigenvalue spectrum of the covariance matrix $\mathbf{X}^{\mathrm{rand}} = \frac{1}{N}(\mathbf{W}^{\mathrm{rand}})^{\top}\mathbf{W}^{\mathrm{rand}}$ 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 look similar, only anti-grokking exhibits Correlation Traps.

Figure 1 shows the phenomenon in three controlled settings: an MNIST MNIST–MLP, a transformer trained on Modular Addition (MODADD), and a GPT2 transformer trained for long-horizon grokking (Wang et al., 2024). 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 3 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 this overfitting phase in a well-studied grokking setting where standard train/test curves alone would not distinguish it from pre-grokking. Detection is only the first step: a separated mode can be benign or task-relevant, so we evaluate output-space activity with random-probe JSD and causal relevance with matched interventions.

Our Contributions.

- We identify and characterize anti-grokking in three long-horizon settings and show that it is spectrally distinct from pre-grokking.
- We define a reproducible, weight-only shuffled-spectrum diagnostic using the public `WeightWatcher` v0.7.5.5 MP/TW implementation, and directly check its fitted null by phase.
- We rule out several simpler explanations with repeated shuffles, layer-wise centering, global-norm comparisons, and matched largest-weight edits.
- We validate behavioral relevance with a label-free JSD ranking test and a continued-training intervention in which targeted trap removal outperforms a correction-budget-matched scalar edit.
- We report exploratory GPT-OSS 20B/120B trap profiles as a scale-up observation, without treating trap presence alone as evidence of harmful overfitting.

Most importantly, the resulting workflow separates three questions that are often conflated: the shuffled spectrum detects candidate modes, random-probe JSD ranks their output influence, and labeled ablation determines the sign of their task effect.

2 Related Work

Grokking and long-horizon generalization. Grokking was introduced by Power et al. (Power et al., 2022) 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 (Liu et al., 2022; Nanda et al., 2023; Lee et al., 2024; Wang et al., 2024). 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 training.

Random-matrix diagnostics of neural-network weights. Our diagnostic builds on spectral approaches to neural-network weight matrices, especially Random Matrix Theory (RMT) and the `WeightWatcher` framework (Martin, 2018-2024; Martin and Hinrichs, 2025; Martin and Mahoney, 2021; Martin et al., 2021). 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 spectrum after entry-wise randomization and track separated MP/TW-edge modes through extended training. Correlation Traps were proposed in SETOL (Martin and Hinrichs, 2025); the contribution here is the discovery and controlled characterization of the anti-grokking phase through these modes, together with fitted-null checks, marginal controls, a data-free JSD ranking test, and a continued-training intervention.

Self-averaging and overfitting. Our interpretation connects to statistical-mechanics accounts of glassy learning, where poor generalization can reflect sample-specific structure rather than a single stable rule (Hopfield, 1982; Amit et al., 1985; Gardner, 1988; Seung et al., 1992; Bös, 1998). The MP law supplies a baseline for the randomized spectrum. A separated edge eigenvalue is evidence that this particular spectral observable departs from that baseline. In a tractable spiked-regression model, Li and Sonthalia show that generalization risk depends jointly on spike strength and target–spike alignment, producing benign, tempered, or catastrophic overfitting regimes (Li and Sonthalia, 2025). We use this result as the theoretical basis for the empirical analysis below and the formal results in the appendix. MP/TW identifies the atypical eigenvalues/vectors, while the JSD and direct ablation we later show tests its influence on model outputs.

3 WeightWatcher, Random Matrix Theory, and Correlation Traps

The open-source `WeightWatcher` tool (`WeightWatcher`) implements random-matrix analyses of neural-network layers Martin (2018-2024). We use public version v0.7.5.5 rather than a hand-fitted MP curve. For each randomized ESD, `WeightWatcher` rescales the spectrum, estimates the MP variance at the observed aspect ratio using kernel density estimation, fits the bulk, detects left- and right-edge outliers when the fitted $\widehat{\sigma} < 1$, reports randomized scale/soft-rank/bulk/spike statistics and fit plots, and applies its finite-size TW edge correction. Fitting, thresholding, and finite-size correction are therefore fixed by the cited implementation rather than tuned separately for each phase.

3.1 The Marchenko-Pastur (MP) self-averaging baseline

For a layer weight matrix $\mathbf{W} \in \mathbb{R}^{N\times M}$, define the layer covariance

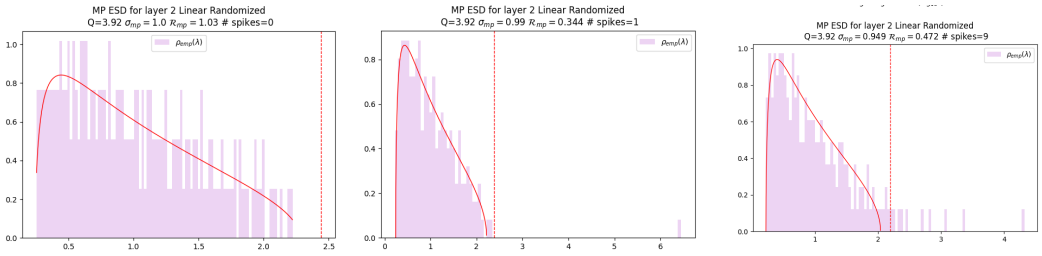
$$\mathbf{X} := \frac{1}{N}\mathbf{W}^\top\mathbf{W} \in \mathbb{R}^{M\times M}. \tag{1}$$

Let $\{\lambda_i\}_{i=1}^M$ be the eigenvalues of $\mathbf{X}$. The empirical spectral density (ESD) is

$$\rho_{\mathrm{emp}}(\lambda) := \frac{1}{M}\sum_{i=1}^{M}\delta(\lambda-\lambda_i). \tag{2}$$

If the 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, $\rho_{\mathrm{emp}}(\lambda)$ converges to the Marchenko–Pastur density Marchenko and Pastur (1967).

$$\rho_{\mathrm{MP}}(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{\lambda}\mathbf{1}_{[\lambda_-,\lambda_+]}(\lambda), \tag{3}$$



(a) (b) (c)

Figure 2: **Randomized ESDs, MP fits, and emergence of Correlation Traps.** (a) Example of the ESD of $\mathbf{X}^{\mathrm{rand}}$ compared with a Marchenko–Pastur (MP) fit. In a self-averaging randomized layer, the spectrum is described by the MP bulk and no large right-edge outliers remain. (b,c) Examples of randomized spectra from Layer 2 of the MNIST–MLP. Right before collapse, the randomized spectrum already shows a dominant spike near $\lambda_{\mathrm{trap}} \approx 10^{6.5}$; during the final anti-grokking stage, multiple spikes lie well beyond the MP bulk. We call these outliers Correlation Traps.

with edges

$$\lambda_{\pm} = \sigma^2 \left(1 \pm Q^{-1/2}\right)^2. \tag{4}$$

At finite N, the right edge λ_+ lives within the scale of the Tracy-Widom (TW) fluctuations.

Also, and importantly, the MP singular and/or eigenvectors $\mathbf{v}$ are delocalized with randomly distributed components. That is, we may expect $|v_j|^2 \sim \frac{1}{M}$ up to fluctuations.

The MP bulk is a **self-averaging** spectral baseline. In the standard large-N model, the ESD concentrates onto the deterministic MP law,

$$\rho_{\mathrm{emp}}(\lambda) \;\to\; \rho_{\mathrm{MP}}(\lambda), \quad N, M \to \infty,\; Q = \tfrac{N}{M} \text{ fixed}.$$

A separated edge mode shows that this particular spectral observable is not described by the fitted bulk. Whether another observable fails to concentrate depends on its overlap with the exceptional direction. This qualification is essential: an outlier $(\lambda_{trap}, \mathbf{v}_{trap})$ is a candidate non-self-averaging structure, not an automatic certificate of harmful overfitting.

3.2 Entry-wise shuffling and Correlation Traps

We randomize the weight matrix element-wise and form

$$\mathbf{W} \mapsto \mathbf{W}^{\mathrm{rand}}. \tag{5}$$

Entry-wise randomization is a permutation without replacement. We use MP/TW as a conditional empirical null and report fitted-MP goodness-of-fit by phase and repeat the randomization across fresh shuffles. Under an adequate null, the ESD of $\mathbf{X}^{\mathrm{rand}} = \frac{1}{N}(\mathbf{W}^{\mathrm{rand}})^\top \mathbf{W}^{\mathrm{rand}}$ should follow an MP bulk and its bulk eigenvectors should be delocalized.[1]

Laplace-like weights do not by themselves invalidate this null. A Laplace law is heavier-tailed than a Gaussian but has finite moments and remains in a standard finite-moment MP/TW regime. Luqman et al. fit Gaussian and Laplace models by maximum likelihood and compare them with Kolmogorov–Smirnov statistics across pretrained and fine-tuned CNN tensors, including ResNet-18/20/50, Inception-v3, MobileNet-V2, and VGG16 (Luqman et al., 2025). Their results support treating Gaussian-like and Laplace-like weights as realistic finite-moment cases; Martin and Mahoney summarize the relevant RMT universality classes (Martin and Mahoney, 2019, Table 1).

We note that, in our experiments, shuffled-spectral outliers are a more reliable signature of harmful overfitting than simpler marginal weight statistics. Although elementary statistics may be more familiar, they are not necessarily simpler as diagnostics. Our choice rests on three features: consistency across our experiments, theoretically grounded outlier thresholds, and an explicit statistical null model.

[1] In the standard random-matrix model, bulk vector components follow Porter–Thomas statistics Porter and Thomas (1956); Potters and Bouchaud (2021).

The spectral diagnostic tracks overfitting in both the language-model and modular-addition experiments, whereas marginal kurtosis tracks it only in the former. The Marchenko–Pastur edge supplies a spectral boundary under the null assumptions, and finite-size corrections calibrate deviations from that boundary. Marginal moments, including the mean, variance, and kurtosis, have no established cross-task threshold for harmful overfitting. The random-matrix null also gives spectral deviations a defined statistical interpretation without requiring a parametric fit to the marginal weight tails. Such fits would introduce additional choices of distribution and fitting range, without guaranteeing that the weights exhibit a meaningful power-law regime.

Figure 2 shows this diagnostic. In a well-behaved layer, the randomized ESD follows the fitted MP bulk and no large right-edge outliers appear; see Fig. 2(a). During anti-grokking, however, the randomized spectrum develops separated spikes well beyond the MP edge; see Fig. 2(b,c). These spikes are structural outliers that survive entry-wise shuffling (and frequently localize).

We call these outliers **Correlation Traps**. If λ_{trap} is an eigenvalue of $\mathbf{X}^{\mathrm{rand}}$ and λ_+^{MP} is the fitted MP right edge, then a trap is an eigenvalue that lies significantly beyond the right MP edge:

$$\text{Correlation Trap:} \qquad \lambda_{trap} > \lambda_+^{\mathrm{MP}} + \Delta_{\mathrm{TW}} \tag{6}$$

where Δ_{TW} denotes the scale of finite-size Tracy-Widom fluctuations.

By the Baik–Ben Arous–Péché (BBP) theory, an eigenvalue that separates beyond the ordinary edge-fluctuation scale should be treated as a structural outlier rather than a bulk fluctuation Baik et al. (2005). This motivates the fitted MP/TW threshold. It does not supply a direct classification of these outlier Traps : Li and Sonthalia's spiked-risk theory makes clear that spectral strength and task overlap jointly determine whether a trap is benign, tempered, or catastrophic (Li and Sonthalia, 2025).

Our setting is nonparametric and implemented as a practical diagnostic. Some traps are behaviorally inert, some support useful computation, and some are associated with harmful overfitting. In order to classify the effect we introduce a data-free JSD score (Section 5) for ranking output influence; and a labeled ablation is done to determine whether the trap helps or hurts test performance.

Table 1: **Trap-detection algorithm.** Full procedure 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	Permute the entries W_{ij} without replacement to form $\mathbf{W}^{\mathrm{rand}}$, retaining the permutation for any later intervention.
3	Run SVD($\mathbf{W}^{\mathrm{rand}}$), form the eigenvalue density, $\rho_{emp}(\lambda)$ (or ESD)
4	Fit the ESD $\rho_{emp}(\lambda)$ to the Marchenko–Pastur (MP) law with `WeightWatcher` v0.7.5.5 at the observed aspect ratio.
5	Identify eigenvalues $\lambda_{trap} > \lambda_+^{\mathrm{MP}} + \Delta_{\mathrm{TW}}$; these are Correlation Traps

The `WeightWatcher` tool detects Correlation Traps, as outlined in Table 1. Table 2 provides a direct phase-level check of the null for FC1 in the no-weight-decay MLP. Pre-grokking and peak generalization have small KS distances and no traps. At collapse, the fitted-MP distance increases by more than an order of magnitude and nine traps appear. Because the MP parameters are fitted and the permuted entries are dependent, the nominal KS p-value is descriptive rather than a full hypothesis test; important evidence to consider is the distance, phase contrast, and repeated-shuffle stability. A paired with-replacement control yields essentially identical phase separation and checkpoint-level trap counts (Pearson $r = 0.9993$, Spearman $\rho = 0.9519$). We retain permutation without replacement because it preserves the complete scalar-weight multiset exactly; Appendix F.1 gives the full comparison.

Conditional risk classification. Appendix A gives a self-contained derivation for a calibrated rank-one Gaussian surrogate and shows that spectral detectability and risk classification have distinct thresholds, with target-spike overlap controlling whether a separated mode raises or lowers risk. The core idea is that MP/TW detects candidate traps, random-probe JSD behaviourally charecterizes them , and labeled ablation confirms their effect and the predictive capability of the random-probe JSD.

Table 2: **FC1 randomized-spectrum check, `WD=0`.** Fitted MP scale, KS distance to the fitted MP CDF, nominal KS p-value, and trap count at one checkpoint from each phase.

FC1 state	$\widehat{\sigma}_{\mathrm{MP}}$	**KS** D	**nominal KS** p	**traps**
Pre-grokking	≈ 1.002	0.0120	≈ 1.0	0
Grokking (maximum test accuracy)	≈ 0.999	0.0212	≈ 1.0	0
Anti-grokking (collapse)	≈ 0.949	0.3044	1.877×10^{-5}	9

4 Empirical Results: Correlation Traps Track Anti-Grokking

We study three standard grokking benchmarks, denoted MNIST–MLP , ModAdd, and GPT2–Compose. The first, MNIST–MLP, is a 3-layer 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 $M = 200$ in each hidden layer and is trained with AdamW, MSE loss on one-hot targets, and learning rate $\mathrm{lr} = 5 \times 10^{-4}$ for up to 10^7 optimization steps. The main runs use `WD=0` plus a control with `WD=0.01` which measures the effect of weight decay. Full hyperparameters appear in Appendix B.

The second benchmark, ModAdd, is a small one-layer transformer trained on the modular-addition task $\mathbf{x} + \mathbf{y} \bmod P$. We again train far beyond the point at which the model first generalizes. Architecture details, phase summaries, and per-layer trap counts appear in Appendix B.3. Pre-grokking, grokking, and anti-grokking describe the observed train/test trajectories.

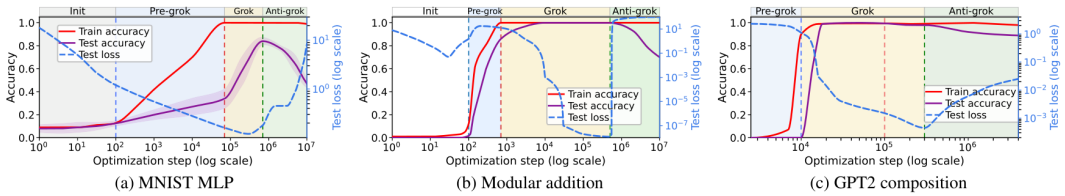
The third benchmark, GPT2–Compose, follows the synthetic composition task of GrokkedTransformer (Wang et al., 2024). It uses a synthetic knowledge graph with atomic one-hop facts (h, r, t), queried as $(h, r) \mapsto t$, and latent composition rules that produce inferred two-hop facts. For instance, atomic facts (h, r_1, m) and (m, r_2, t) imply the inferred query $(h, r_1, r_2) \mapsto t$. Following the original setup, atomic facts are split into $\mathrm{atomic}_{\mathrm{ID}}$ and $\mathrm{atomic}_{\mathrm{OOD}}$. Training includes all atomic facts and a random subset of inferred facts derived only from $\mathrm{atomic}_{\mathrm{ID}}$, denoted $\mathrm{train_inferred}_{\mathrm{ID}}$; $\mathrm{test_inferred}_{\mathrm{ID}}$ is held out. Thus, "test" measures in-distribution compositional generalization over unseen inferred facts from the same atomic pool. Appendix B.4 gives the full construction.

The supplemental material (See Appendix H) contains the corresponding training and analysis code.

4.1 MLP: A third phase after grokking

Following Liu et al. (Liu et al., 2023), our standard setup scales the default parameters by $\alpha_{\mathrm{init}} = 8$ to induce grokking on MNIST. We also test alternative values of α_{init}, which shift the timing of grokking and anti-grokking; the phase-aligned trap signal persists across these timing shifts (Appendix D).

Figure 1(a) shows the full long-horizon trajectory for the MNIST–MLP trained on the MNIST subset. The run exhibits the familiar pre-grokking $\rightarrow$ 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. Likewise, as shown in Fig. 3, the test loss reaches a minimum at some point. This is the regime (green) we call anti-grokking.



(a) MNIST MLP (b) Modular addition (c) GPT2 composition

Figure 3: **Grokking dynamics and test loss.** Train accuracy (red), test accuracy (purple), and test loss (blue) are shown for (a) an MLP on MNIST, (b) Modular Addition (ModAdd), and (c) GPT2. Phase boundaries are determined separately for each trajectory from its held-out loss/accuracy history.

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, the average number of detected traps is effectively zero. During grokking, trap count remains near zero. It is only when late-stage anti-grokking begins that trap count rises sharply. The controls below test whether this separation can be reduced to mean, scale, or a few largest coordinates.

Table 5 makes the phase separation explicit. In both the no-weight-decay setting (`WD=0`) and the weight-decay control (`WD>0`), traps are absent during pre-grokking and grokking and appear only in anti-grokking. The no-weight-decay run exhibits both a larger trap count and a more severe drop in test accuracy, suggesting that trap count reflects not only onset but also severity.

4.2 GPT2: PHASE-ALIGNED TRAP GROWTH IN A TRANSFORMER

We also train the GrokkedTransformer GPT2 model long enough to expose the anti-grokking regime (Wang et al., 2024). Figure 1(c) shows training accuracy, held-out accuracy, and per-layer trap counts. The model first memorizes, then generalizes, and later loses a substantial fraction of its held-out accuracy while retaining nearly perfect training accuracy. Trap counts become nonzero around the transition out of the minimum-test-loss region shown in Fig. 3(c). This supplies a second transformer setting in which the trap count follows phase.

4.3 WEIGHT DECAY SUPPRESSES ANTI-GROKKING (TRAP GROWTH & OVERFITTING)

A useful control is the MNIST–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 B.2; Table 5). Thus the late-stage phase persists under regularization, but both its shuffled-spectrum signature and its generalization loss are weaker. The scale and centering controls below are needed before giving that association a structural interpretation.

4.4 ADDITIONAL VALIDATION: PERTURBATION STUDY

Appendix D repeats the MLP analysis across initialization multipliers α_{init} to test whether trap counts merely track weight scale. Although changing α_{init} shifts global norms and phase timing, trap counts remain near zero in pre-grokking and grokking and become nonzero only in anti-grokking.

4.5 CENTERING AND SIMPLER MARGINAL CONTROLS

At the analyzed anti-grokking checkpoint, the empirical mean (standard deviation) is -0.1316 (1.0549), -0.1778 (0.2757), and -0.00478 (0.2522) for linear layers 1, 3, and 5. Because a nonzero mean can generate a rank-one contribution, we repeat the full shuffled analysis after subtracting each layer's empirical mean. We also compare with two largest-weight edits. In each layer, K is the number of detected modes; the matched scalar edit changes the K largest-magnitude weights using the same Frobenius correction budget as the trap edit.

Table 3: **Centering and largest-weight controls.** Total shuffled outliers at the same anti-grokking checkpoint over 50 fresh entry permutations.

One-shot condition	**Total shuffled outliers, mean $\pm$ SD**
Unmodified	38.2 ± 2.4
Subtract each layer's empirical mean	36.4 ± 1.8
Largest-$\lvert w\rvert$, same K and $\lVert \Delta W \rVert_F$	37.6 ± 2.4
Hard-zero the same K largest-$\lvert w\rvert$ entries	22.8 ± 3.6

Neither a rank-one mean contribution nor the same K largest coordinates explains most of the signal. Hard-zeroing those coordinates removes some modes, as expected, but leaves more than half of the outliers. Appendix D provides the complementary phase-level scale comparison: global ℓ_2 norm is 17.52 ± 0.00, 16.80 ± 0.22, and 27.64 ± 8.39 in pre-grokking, grokking, and anti-grokking, while mean trap count is 0.00, 0.00, and 4.56 ± 3.03. The norm changes before collapse; trap count remains zero until generalization is lost. This does not establish independence from every marginal statistic, but it rules out a simple global-norm threshold.

5 RANKING TRAP INFLUENCE WITH DATA-FREE JSD

The randomized spectrum detects candidate modes, but it does not determine whether a mode affects the model's outputs. We therefore define a data-free JSD diagnostic ablation test. For each included two-dimensional matrix, we save the entry permutation used to form $\mathbf{W}^{\mathrm{rand}}$. After detecting a separated shuffled singular mode, we replace or clip that mode, reconstruct the randomized matrix, invert the saved permutation, and write the intervention back in the original layer coordinates. Figure 8 uses a matched Gaussian replacement for the mapped FC1 direction.

Let M_θ denote the original model and $M_{\theta\setminus k}$ the model after replacing trap k. Define the temperature-scaled output distributions

$$p_\theta(\mathbf{x};T) = \mathrm{softmax}(z_\theta(\mathbf{x})/T)\,, \qquad p_{\theta\setminus k}(\mathbf{x};T) = \mathrm{softmax}\big(z_{\theta\setminus k}(\mathbf{x})/T\big)\,.$$

Given probe inputs $\mathbf{x}_p$, define the trap-removal score

$$J_k(T) = \mathbb{E}_{\mathbf{x}_p}\left[D_{\mathrm{JS}}\big(p_\theta(\mathbf{x}_p;T)\,\big\|\,p_{\theta\setminus k}(\mathbf{x}_p;T)\big)\right]. \tag{7}$$

Here $z_\theta(\mathbf{x})$ and $z_{\theta\setminus k}(\mathbf{x})$ are the original and trap-removed logits, and D_{JS} denotes the Jensen–Shannon divergence (a symmetric, smoothed version of the KL divergence).

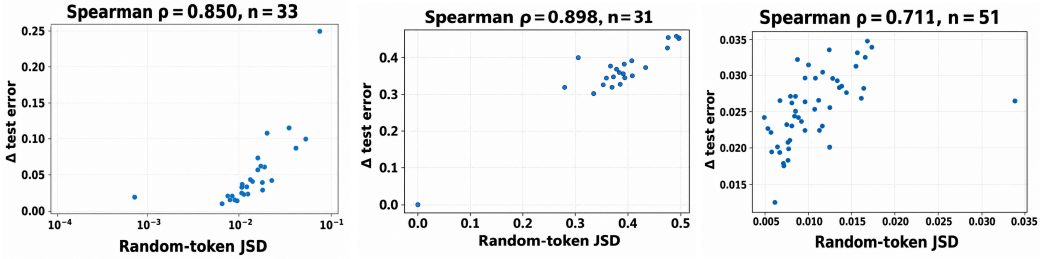
For modular addition and GPT-style models, the probe distribution is random-token sequences. For the MLP, probes are Gaussian inputs matched to the empirical input mean and variance. We use $T = 1$ throughout. A large $J_k(1)$ means that removing the trap changes the model outputs.

At the final `WD=0` MLP checkpoint, we also compare Gaussian-probe JSD with JSD on the official MNIST test set for the same 36 one-trap edits across all detected layers. The two scores are strongly correlated (Spearman $\rho = 0.825$, Pearson $r = 0.859$; both $p < 10^{-4}$), supporting random-probe JSD as a data-free activity ranking. Appendix C.2 gives the paired protocol and full statistics.

JSD measures activity, not sign. When labels are available, the test-error change after the same intervention separates three cases:

- **Harmful trap:** removal reduces test error.
- **Useful trap:** removal increases test error.
- **Behaviorally inert trap:** removal has negligible effect on the tested outputs.

Without labels, JSD can rank modes for follow-up but cannot determine whether their contribution is helpful or harmful.



(a) MNIST–MLP (b) MODADD (c) GPT2–COMPOSE

Figure 4: **JSD diagnostic ablation test for selected MNIST–MLP, MODADD, and GPT2–COMPOSE anti-grokking checkpoints.** Trap-removal score $J_k(T = 1)$ versus measured test-error change for individual modes. Spearman correlations are 0.850 (MLP, $n = 33$), 0.898 (MA, $n = 31$), and 0.711 (GPT2, $n = 51$). JSD is interpreted within a model because its absolute scale depends on architecture and probe distribution.

Figure 4 validates random-probe JSD against the task-based consequence of the same intervention. For MNIST–MLP and MODADD, we evaluate the first, middle, and final anti-grokking checkpoints. For GPT2–COMPOSE, we analyze a representative layer from the first anti-grokking checkpoint (1.5×10^6 steps). Within each model family, the ranking is strong. The result supports JSD as a data-free prioritization score for behaviorally consequential traps; it does not turn a weight-only statistic into a signed generalization metric.

Continued-training intervention. Continuing optimization while removing detected traps recovers held-out accuracy from 50.33% to 64.46%, compared with 50.75% under ordinary continued training and 58.92% under a correction-budget-matched largest-weight edit. Trap removal also restores 100% training accuracy and leaves far fewer residual outliers. Appendix C.4 gives the full protocol and results.

6 Broader implications for frontier-scale models

To test whether the spectral phenotype is confined to small grokking models, we applied the same shuffled-spectrum detector to two open-weight reasoning models released by OpenAI, `gpt-oss-20b` and `gpt-oss-120b` OpenAI (2025a;b). Figure 5 reports the number of detected Correlation Traps per layer.

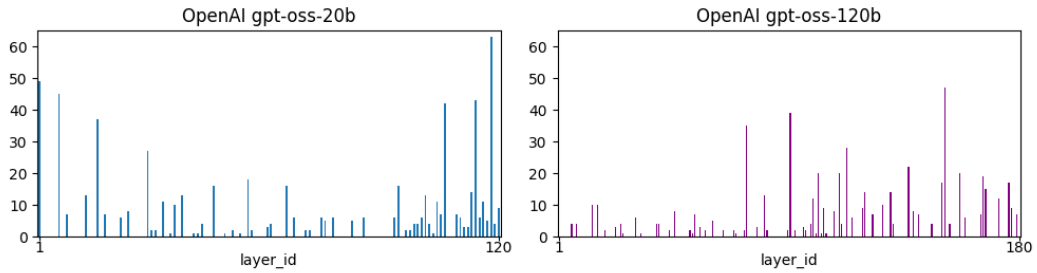


Figure 5: **Exploratory layer-wise Correlation Trap profiles in open-weight LLMs.** Number of detected shuffled-spectrum outliers per layer for OpenAI `gpt-oss-20b` and `gpt-oss-120b`.

The result shows that the same shuffled-spectrum anomaly can occur at much larger scale. It does *not* show that these models are overfit or unsafe: there is no checkpoint trajectory, label-free JSD calibration, or matched intervention for the GPT-OSS models in this study. The practical implication that one can detect layers for follow-up behavioral testing when only weights are available.

7 Conclusion

Grokking is not always the end of training. In the three trajectories studied here, a model can generalize and then lose that generalization under continued optimization while retaining high training accuracy. We call this post-generalization failure mode *anti-grokking*. It is marked by the appearance of separated modes in randomized layer spectra–Correlation Traps–that are absent or nearly absent in pre-grokking and successful grokking.

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. Interestingly during continued optimization, targeted removal produces higher test accuracy and fewer residual outliers than both no intervention and the matched scalar edit.

Correlation Traps provide a practical diagnostic for potentially harmful overfitting even when only the model weights are available. The diagnostic has two stages: identify candidate spectral anomalies from the weights, then rank their output influence using data-free random probes. Random-probe JSD requires neither training nor test data and correlates strongly with measured test-error change in our experiments, supporting its use to prioritize anomalies for further investigation.

Most notably, our GPT-OSS results show that Correlation Traps occur in foundation scale LLMs, sometimes in great numbers. And while it is generally known that there is overfitting in foundation scale LLMs, these results suggest that perhaps not all overfitting is benign (a frequent belief). With behavioral charecterization via JSD left to future work. We hope Correlation Traps can be a useful tool to help improve these and other AI models.

AI USE STATEMENT

In this work, we used generative AI tools for help to draft this paper, in particular to correct typos and to make some paragraphs more concise to fit the 9 page limit. Additionally, we used generative AI tools for refining and verifying proofs , using LEAN. We have reviewed all AI-assisted work such as the lean outputs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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.

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.

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.

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.

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.

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.

Jiping Li and Rishi Sonthalia. Risk phase transitions in spiked regression: Alignment driven benign and catastrophic overfitting. *arXiv preprint arXiv:2510.01414*, 2025.

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.

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=zDiHoIWa0q1.

Ahmed Luqman, Khuzemah Qazi, Murray Patterson, Malik Jahan Khan, and Imdadullah Khan. A data-free analytical quantization scheme for deep learning models, 2025. URL https://arxiv.org/abs/2412.07391.

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

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.

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.

Charles H. Martin and Michael W. Mahoney. Traditional and heavy-tailed self regularization in neural network models, 2019. URL https://arxiv.org/abs/1901.08276.

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.

Charles H. Martin, Tongsu (Serena) 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.

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.

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

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

C. E. Porter and R. G. Thomas. Fluctuations of nuclear reaction widths. *Physical Review*, 104(2): 483–491, 1956.

Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists*. Cambridge University Press, 2021. ISBN 9781108768900.

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.

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.

Dietrich von Rosen. Moments for the inverted wishart distribution. *Scandinavian Journal of Statistics*, 15(2):97–109, 1988.

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.

A RIGOROUS DERIVATION OF THE TRAP–RISK CONNECTION

This appendix proves the risk formula and phase statements summarized in the main text. The proof is self-contained for the calibrated Gaussian surrogate.

A.1 RISK OF THE RANK-ONE GAUSSIAN SURROGATE

For each $d > n$, let $x_i \in \mathbb{R}^d$ be independent with

$$x_i \sim \mathcal{N}(0, \Sigma), \qquad \Sigma = \tau^2(I_d + \gamma u u^\top), \qquad \|u\| = 1,$$

and let $y_i = \alpha x_i^\top \beta + \varepsilon_i$, where the ε_i are independent of the design, centered, and have variance τ_ε^2. Write

$$X = [x_1, \dots, x_n], \qquad \widehat{b} = X(X^\top X)^{-1} y, \qquad \rho = \beta^\top u, \qquad b_\perp^2 = \|\beta\|^2 - \rho^2.$$

The inverse exists almost surely. For an independent test observation (x_0, y_0), define excess risk by

$$\mathcal{R}_{d,n} = \mathbb{E}[(y_0 - x_0^\top \widehat{b})^2] - \tau_\varepsilon^2.$$

Theorem A.1 (Gaussian rank-one risk)**.** *Fix $c > 1$, $\gamma \geq 0$, $\alpha, \tau > 0$, and $\tau_\varepsilon^2 < \infty$. Suppose $d, n \to \infty$ with $d/n \to c$, $\beta^\top u \to \rho$, and $\|\beta\|^2 - (\beta^\top u)^2 \to b_\perp^2 < \infty$. Then*

$$\mathcal{R}_{d,n} \longrightarrow \mathcal{R}_c(\gamma) = \frac{\tau_\varepsilon^2}{c-1} + \alpha^2\tau^2\left(1 - \frac{1}{c}\right)\left[b_\perp^2 + \rho^2\frac{c^2(1+\gamma)}{(c+\gamma)^2}\right]. \tag{8}$$

Proof. By rotational invariance, take $u = e_1$ and put $D = I_d + \gamma e_1 e_1^\top$. Standardize the design as

$$Y := X/\tau = \begin{bmatrix} \sqrt{1+\gamma}\, v^\top \\ Z \end{bmatrix}, \qquad v \sim \mathcal{N}(0, I_n), \qquad Z \in \mathbb{R}^{m \times n}, \quad m = d - 1, \tag{9}$$

where v and Z are independent and the entries of Z are independent standard Gaussians. Define

$$B = Z^\top Z, \qquad H = Y^\top Y = B + (1+\gamma)vv^\top, \qquad P = YH^{-1}Y^\top, \qquad Q = I_d - P.$$

Here P and Q are complementary orthogonal projectors. The coefficient error is

$$\widehat{b} - \alpha\beta = -\alpha Q\beta + \tau^{-1} Y H^{-1}\varepsilon. \tag{10}$$

Since the training noise is centered and independent, the cross term vanishes. Conditional averaging over the test feature and then the training noise therefore gives

$$\mathcal{R}_{d,n} = \alpha^2\tau^2\mathbb{E}[\beta^\top Q D Q \beta] + \tau_\varepsilon^2 \mathbb{E}\,\mathrm{tr}(H^{-1}Y^\top D Y H^{-1}). \tag{11}$$

We first record the inverse-Gram limits. Since $B \sim W_n(m, I_n)$, standard inverse-Wishart moments (von Rosen, 1988) give, for $m > n + 3$,

$$\mathbb{E}\,\mathrm{tr}(B^{-1}) = \frac{n}{m-n-1}, \tag{12}$$

$$\mathbb{E}\,\mathrm{tr}(B^{-2}) = \frac{n(m-1)}{(m-n)(m-n-1)(m-n-3)}, \tag{13}$$

$$\mathrm{Var}(\mathrm{tr}\, B^{-1}) = \frac{2n(m-1)}{(m-n)(m-n-1)^2(m-n-3)}. \tag{14}$$

Thus

$$\mathrm{tr}\, B^{-1} \longrightarrow \frac{1}{c-1} \quad \text{in } L^2, \qquad \mathbb{E}\,\mathrm{tr}\, B^{-2} = O(n^{-1}). \tag{15}$$

Let $t = v^\top B^{-1} v$. Conditional Gaussian quadratic-form identities give

$$\mathbb{E}[t \mid B] = \mathrm{tr}\, B^{-1}, \qquad \mathrm{Var}(t \mid B) = 2\,\mathrm{tr}\, B^{-2}.$$

Equations (14)– (15) imply

$$t \longrightarrow \frac{1}{c-1} \quad \text{in } L^2. \tag{16}$$

We now evaluate the bias. Sherman–Morrison gives

$$H^{-1} = B^{-1} - \frac{(1+\gamma)B^{-1}vv^\top B^{-1}}{1+(1+\gamma)t}. \tag{17}$$

The spike-coordinate entry of P and its complementary entry in Q are therefore

$$P_{11} = \frac{(1+\gamma)t}{1+(1+\gamma)t}, \qquad Q_{11} = \frac{1}{1+(1+\gamma)t} \longrightarrow a := \frac{c-1}{c+\gamma}.$$

Because $0 \le Q_{11} \le 1$, its first and second moments converge to a and a^2. Write $\beta = \rho e_1 + \beta_\perp$. Rotational invariance in the m bulk coordinates, $\operatorname{tr} Q = d-n$, and $\operatorname{tr} Q_{bb} = d-n-Q_{11}$ imply

$$\mathbb{E}[\beta_\perp^\top Q \beta_\perp] = \frac{\|\beta_\perp\|^2}{m}\mathbb{E}\operatorname{tr} Q_{bb} \longrightarrow \left(1-\frac{1}{c}\right) b_\perp^2. \tag{18}$$

The mixed first moment is zero by the same rotational symmetry. Moreover, since $Q^2 = Q$,

$$\mathbb{E}|e_1^\top Q\beta_\perp|^2 = \frac{\|\beta_\perp\|^2}{m}\mathbb{E}\sum_{j=2}^{d} Q_{1j}^2 = \frac{\|\beta_\perp\|^2}{m}\mathbb{E}(Q_{11}-Q_{11}^2) \longrightarrow 0.$$

It follows that

$$\mathbb{E}[\beta^\top Q\beta] \longrightarrow \left(1-\frac{1}{c}\right) b_\perp^2 + \rho^2 a, \tag{19}$$

$$\mathbb{E}[(e_1^\top Q\beta)^2] \longrightarrow \rho^2 a^2. \tag{20}$$

Using $QDQ = Q + \gamma Q e_1 e_1^\top Q$, the bias term in (11) converges to

$$\alpha^2\tau^2\left[\left(1-\frac{1}{c}\right)b_\perp^2 + \rho^2(a+\gamma a^2)\right] = \alpha^2\tau^2\left(1-\frac{1}{c}\right)\left[b_\perp^2 + \rho^2\frac{c^2(1+\gamma)}{(c+\gamma)^2}\right], \tag{21}$$

because $a+\gamma a^2 = c(c-1)(1+\gamma)/(c+\gamma)^2$.

It remains to evaluate the training-noise term. Since

$$Y^\top DY = H + \gamma(1+\gamma)vv^\top,$$

that term, divided by τ_ε^2, equals

$$\mathbb{E}\operatorname{tr} H^{-1} + \gamma(1+\gamma)\mathbb{E}[v^\top H^{-2} v]. \tag{22}$$

Equation (17) gives

$$0 \le \operatorname{tr}(B^{-1}-H^{-1}) = \frac{(1+\gamma)v^\top B^{-2}v}{1+(1+\gamma)t} \le (1+\gamma)v^\top B^{-2}v,$$

$$v^\top H^{-2}v = \frac{v^\top B^{-2}v}{[1+(1+\gamma)t]^2} \le v^\top B^{-2}v.$$

Conditional expectation over v turns the final quadratic form into $\operatorname{tr} B^{-2}$. Hence (15) implies

$$\mathbb{E}\operatorname{tr} H^{-1} \longrightarrow \frac{1}{c-1}, \qquad \mathbb{E}[v^\top H^{-2}v] \longrightarrow 0.$$

Substitution into (22) shows that the noise contribution tends to $\tau_\varepsilon^2/(c-1)$. Combining this with (21) proves (8). □

Theorem A.1 is the Gaussian specialization of the well-specified formula in Li and Sonthalia (2025, Theorem 1). The direct proof is included to make every assumption needed by the above explicit.

A.2 Risk sign, spectral detection, and phase limits

Define

$$K_c = \alpha^2\tau^2\left(1-\frac{1}{c}\right), \qquad h_c(\gamma) = \frac{c^2(1+\gamma)}{(c+\gamma)^2}.$$

Corollary A.2 (Finite-c sign and the detectable harmful window)**.** *For $c > 1$, $\gamma > 0$, and $\rho \neq 0$,*

$$\mathcal{R}_c(\gamma) - \mathcal{R}_c(0) = K_c\rho^2\frac{\gamma[c(c-2)-\gamma]}{(c+\gamma)^2}. \tag{23}$$

Thus the spike raises risk relative to the spike-free surrogate exactly when $\gamma < c(c-2)$, leaves it unchanged at equality, and lowers it exactly when $\gamma > c(c-2)$.

For the Gaussian covariance surrogate, the population spike is separated from the sample-covariance bulk when $\gamma > \sqrt{c}$. Consequently, a detectable but risk-increasing interval exists exactly when

$$c > \frac{3+\sqrt{5}}{2}, \qquad \sqrt{c} < \gamma < c(c-2). \tag{24}$$

Proof. Direct algebra gives

$$h_c(\gamma) - 1 = \frac{\gamma[c(c-2)-\gamma]}{(c+\gamma)^2},$$

which proves (23) and its sign. The normalized population covariance eigenvalue is $1+\gamma$. The spiked-covariance transition of Baik et al. (2005, Theorem 1.2) is $1+\gamma > 1+\sqrt{c}$, equivalently $\gamma > \sqrt{c}$; the corresponding sample bulk edge is $(1+\sqrt{c})^2$.

It remains to determine when the two inequalities are compatible. Put $t = \sqrt{c} > 1$. Then

$$c(c-2) > \sqrt{c} \iff t^3 - 2t - 1 > 0 \iff (t+1)(t^2-t-1) > 0 \iff t > \frac{1+\sqrt{5}}{2}.$$

Squaring the last inequality gives $c > (3+\sqrt{5})/2$, proving (24). □

Corollary A.3 (Sequential risk phases)**.** *After taking the proportional limit at fixed c, let $c \to \infty$ and assume $\tau_\varepsilon^2 = O(1)$, $|\rho_c| \to \rho > 0$, and $b_{\perp,c}^2 \to b_\perp^2 < \infty$. Then*

$$\mathcal{R}_c \longrightarrow \begin{cases} \alpha^2\tau^2[b_\perp^2 + \rho^2(1+\gamma_0)], & \gamma_c \to \gamma_0 \in [0,\infty), \\ +\infty, & \gamma_c \to \infty, \quad \gamma_c = o(c^2), \\ \alpha^2\tau^2[b_\perp^2 + \rho^2/\phi], & \gamma_c/c^2 \to \phi \in (0,\infty), \\ \alpha^2\tau^2 b_\perp^2, & \gamma_c/c^2 \to \infty. \end{cases} \tag{25}$$

In particular, the last regime is benign if and only if $b_{\perp,c}^2 \to 0$.

Proof. The observation-noise term tends to zero and $K_c \to \alpha^2\tau^2$. For a bounded convergent spike,

$$h_c(\gamma_c) = \frac{1+\gamma_c}{(1+\gamma_c/c)^2} \longrightarrow 1+\gamma_0.$$

If $\gamma_c \to \infty$ and $\gamma_c = o(c^2)$, then

$$\frac{1}{h_c(\gamma_c)} = \frac{(c+\gamma_c)^2}{c^2(1+\gamma_c)} \leq \frac{1}{\gamma_c} + \frac{2}{c} + \frac{\gamma_c}{c^2} \longrightarrow 0,$$

so $h_c(\gamma_c) \to \infty$. If $\gamma_c/c^2 \to \phi \in (0,\infty)$, the exact representation

$$h_c(\gamma_c) = \frac{c^{-2} + \gamma_c/c^2}{(c^{-1} + \gamma_c/c^2)^2}$$

gives $h_c(\gamma_c) \to 1/\phi$. Finally, if $\gamma_c/c^2 \to \infty$, then

$$0 \leq h_c(\gamma_c) \leq \frac{c^2}{\gamma_c^2} + \frac{c^2}{\gamma_c} \longrightarrow 0.$$

Substitution in Theorem A.1 proves (25). □

Lemma A.4 (Absolute versus target-normalized risk)**.** *The noiseless target energy is*

$$S_c := \mathbb{E}[(\alpha x^\top \beta)^2] = \alpha^2 \tau^2 [b_\perp^2 + (1+\gamma)\rho^2], \tag{26}$$

and

$$0 \le \mathcal{R}_c - \frac{\tau_\varepsilon^2}{c-1} \le \left(1 - \frac{1}{c}\right) S_c. \tag{27}$$

Thus the catastrophic statement in Corollary A.3 concerns absolute excess risk; the target energy may grow at the same time.

Proof. Equation (26) follows from $\beta^\top \Sigma \beta$. Since

$$0 \le h_c(\gamma) = (1+\gamma)\left(\frac{c}{c+\gamma}\right)^2 \le 1+\gamma,$$

the bound follows from (8). □

A.3 Sign determination

The preceding results concern a calibrated statistical surrogate. The next identities apply directly to an original network and its intervened version.

Proposition A.5 (Exact signed-ablation identity)**.** *Let $z(x)$ and $z^-(x) = z(x) + \Delta z(x)$ be the original and intervened logits, and put $p = \mathrm{softmax}(z)$ and $p^- = \mathrm{softmax}(z^-)$. If e_Y is the one-hot label vector, then their expected cross-entropy risks satisfy*

$$R_{\mathrm{CE}}(z^-) - R_{\mathrm{CE}}(z) = \mathbb{E}[(p - e_Y)^\top \Delta z] + \mathbb{E}\,\mathrm{KL}(p \| p^-). \tag{28}$$

Hence the intervention improves cross-entropy exactly when

$$\mathbb{E}[(p - e_Y)^\top \Delta z] < -\mathbb{E}\,\mathrm{KL}(p \| p^-). \tag{29}$$

Proof. Let $A(z) = \log \sum_k e^{z_k}$. Pointwise,

$$\ell_{\mathrm{CE}}(z, Y) = A(z) - e_Y^\top z$$

and

$$\mathrm{KL}(p \| p^-) = A(z^-) - A(z) - p^\top (z^- - z).$$

Adding $(p - e_Y)^\top (z^- - z)$ proves (28); taking expectations and rearranging gives (29). □

Lemma A.6 (Local geometry of random-probe JSD)**.** *For fixed finite logits $z \in \mathbb{R}^K$, let $p = \mathrm{softmax}(z)$ and $F(p) = \mathrm{diag}(p) - pp^\top$. With natural logarithms,*

$$\mathrm{JSD}(p_z, p_{z+\delta z}) = \frac{1}{8}\,\delta z^\top F(p) \delta z + o(\|\delta z\|^2). \tag{30}$$

At temperature $T > 0$, the leading term is multiplied by T^{-2} and p is replaced by $\mathrm{softmax}(z/T)$.

Proof. The softmax differential is $\delta p = F(p)\delta z + O(\|\delta z\|^2)$. Taylor expansion of the two KL terms defining JSD around the interior point p gives

$$\mathrm{JSD}(p, p + \delta p) = \frac{1}{8} \sum_k \frac{\delta p_k^2}{p_k} + o(\|\delta p\|^2).$$

Writing $D_p = \mathrm{diag}(p)$, the identity $F(p) D_p^{-1} F(p) = F(p)$ proves (30). Replacing z and δz by z/T and $\delta z/T$ gives the temperature-scaled statement. □

Together, Theorem A.1 and Proposition A.5 formalize the separation used in the main text: MP/TW supplies spectral detection, the calibrated spiked-regression surrogate supplies a conditional risk mechanism, JSD measures output influence, and labeled intervention confirms the task sign.

B Experimental Setup and Additional Notes: MLP, MA, and GPT2

B.1 MLP Experimental Setup

We train a Multi-Layer Perceptron (MLP) on a subset of the MNIST dataset using the hyperparameters in Table 4. The training subset is constructed by randomly selecting 100 samples from each of the 10 MNIST classes, yielding a balanced dataset of 1,000 unique training points. The primary long-horizon runs were performed on a single NVIDIA Quadro P2000 GPU and the main 10^7-step MNIST–MLP experiment took approximately 11 hours; a considerable fraction of the wall-clock time is due to checkpointing and saving measurements for later analysis.

The detector is applied to all three two-dimensional linear layers (layers 1, 3, and 5 in the saved model). Table 5 isolates FC1 and FC2 for the weight-decay comparison; Table 16 reports all three layers in the complementary scale perturbation.

Table 4: MLP experimental hyperparameters used in the study.

Parameter	**Value**
Network Architecture	Fully Connected MNIST–MLP
Depth	3 Linear layers (Input $\to$ Hidden1 $\to$ Hidden2 $\to$ Output)
Width	200 hidden units per hidden layer
Activation Function	ReLU
Input Layer Size	784 (flattened MNIST image 28×28)
Output Layer Size	10
Weight Initialization	Default PyTorch (Kaiming Uniform for weights), parameters scaled by 8.0
Bias Initialization	Default PyTorch (Uniform), then scaled by 8.0
Dataset	MNIST
Training Points	1,000 (100 per class, stratified random sampling)
Test Points	Standard MNIST test set (10,000 samples)
Batch Size	200
Loss Function	MSE with one-hot encoded targets
Optimizer	AdamW
Learning Rate (LR)	5×10^{-4}
Weight Decay (WD)	0.0 (main results), 0.01 (control experiment)
AdamW β_1	0.9
AdamW β_2	0.999
AdamW ϵ	10^{-8}
Optimization Steps	10^7
Data Type (PyTorch)	`torch.float64`
Random Seed	0
Software Framework	PyTorch
WeightWatcher Version	v0.7.5.5 Martin (2018-2024)

Note on weight decay. The primary results in the main paper were obtained with weight decay explicitly set to 0. This isolates the long-horizon optimization dynamics from explicit norm regularization. Runs with nonzero weight decay (`WD=0.01`) were performed for comparison and exhibit substantially weaker trap growth and milder loss of generalization.

B.2 MLP Control Experiment with Weight Decay

The weight-decay control uses the same architecture, data, and optimizer as the main MNIST–MLP experiment, but with `WD=0.01`. The central observation is that weight decay suppresses trap growth. As Table 5 shows, anti-grokking under `WD>0` still exhibits traps, but fewer than in the `WD=0` case, and the associated degradation in test accuracy is much smaller. This supports the interpretation that weight decay mitigates the structural anomalies detected by the shuffled-spectrum diagnostic.

B.3 Modular Addition (MA) Experiment

We additionally evaluate the shuffled-spectrum diagnostic on a modular-addition task with a small one-layer transformer.

Table 5: **Average number of Correlation Traps in MNIST–MLP with and without Weight Decay (WD).** Trap counts for FC1 and FC2 at representative checkpoints at the right edge of each phase: pre-grokking ($\sim 10^5$ steps), grokking (10^6), and anti-grokking (10^7). Adding weight decay suppresses trap growth and late-stage collapse, but does not eliminate overfitting entirely.

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

Hyper-parameter	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 6: **Modular Addition phase summary.** Train and test accuracy (mean$\pm$std) across sampled checkpoints in each phase.

Phase	Train Acc. (mean$\pm$std)	Test Acc. (mean$\pm$std)
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

Model architecture. Trap counts are effectively zero in pre-grokking and grokking and rise sharply across many layers in anti-grokking. This reproduces the same qualitative signature observed in the MNIST–MLP: trap-free shuffled spectra in the first two phases and numerous shuffled-spectrum outliers during late-stage collapse. All eight two-dimensional matrices in the one-layer transformer are included in Table 7.

Table 7: **Modular Addition trap counts.** Number of shuffled-spectrum outliers detected by `WeightWatcher` in each layer and phase.

Layer	Pre-grok traps	Grok traps	Anti-grok traps
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	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{4.88 \pm 1.55}$

B.4 GPT2 Experimental Setup

We evaluate the shuffled-spectrum diagnostic on the synthetic composition benchmark from Grokked-Transformer (Wang et al., 2024). The task is generated from a random knowledge graph of atomic one-hop facts and latent two-hop composition rules. Table 8 summarizes the dataset construction, Table 9 defines the evaluation splits, and Table 10 lists the training and analysis configuration.

The training set contains all atomic facts, both ID and OOD, together with a random subset of inferred facts deduced only from $\mathrm{atomic}_{\mathrm{ID}}$. Therefore, the primary train/test comparison measures in-distribution compositional generalization: the model must apply the learned latent composition rule to unseen inferred facts drawn from the same atomic pool used to generate the training compositions. We report $\mathrm{test_inferred}_{\mathrm{OOD}}$ separately, since it probes generalization to inferred facts involving the held-out atomic pool rather than the main grokking transition.

For spectral and ablation analyses, we run WeightWatcher on the transformer block matrices and exclude the positional embedding layer `wpe`. This exclusion avoids treating the positional embedding table as a standard learned linear map in the layerwise trap analysis.

Table 8: **GPT2 synthetic composition dataset.** The benchmark is generated from a random knowledge graph with atomic facts (h, r, t), queried as $(h, r) \mapsto t$, and inferred two-hop queries $(h, r_1, r_2) \mapsto t$ induced by pairs of atomic facts (h, r_1, m) and (m, r_2, t).

Dataset parameter	**Value**
Dataset variant	`composition.2000.200.12.6`
Number of entities	2000
Number of relations	200
Atomic outgoing facts per entity	20
Atomic fact form	(h, r, t)
Atomic query form	$(h, r) \mapsto t$
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 9: **GPT2 train/test split definitions.** The main grokking curve treats $\text{train_inferred}_{\text{ID}}$ as the training curve and $\text{test_inferred}_{\text{ID}}$ as the test curve. The OOD inferred split is reserved as a stronger systematicity diagnostic.

Split	**Included in training?**	**Role in evaluation**
id_atomic	Yes	ID atomic memorization/control accuracy
ood_atomic	Yes	OOD atomic memorization/control accuracy
$\text{train_inferred}_{\text{ID}}$	Yes	Main "train" curve for inferred facts
$\text{test_inferred}_{\text{ID}}$	No	Main "test" curve for ID compositional generalization
$\text{test_inferred}_{\text{OOD}}$	No	Stronger systematicity diagnostic

Table 10: **GPT2 model, training, and evaluation configuration.** The model is an 8-layer GPT2-style decoder-only transformer initialized from scratch and trained with the GrokkedTransformer training code.

Hyper-parameter	**Value**
Architecture	GPT2-style decoder-only transformer
Initialization	From scratch
Number of layers	8
Vocabulary	Synthetic entity, relation, and answer-delimiter tokens added
Maximum sequence length	10
Batch size	512
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	0.001
Schedule	Constant schedule with warmup
Precision	Mixed precision
Training steps	1.5×10^6
Checkpointing	Checkpoints saved throughout training
Evaluation protocol	Exact-match generation
Correctness criterion	Generated answer token matches target before closing delimiter
Reported accuracies	id_atomic, ood_atomic, $\text{train_inferred}_{\text{ID}}$, $\text{test_inferred}_{\text{ID}}$, $\text{test_inferred}_{\text{OOD}}$

Table 11: **GPT2 spectral-analysis convention.** WeightWatcher is applied to learned transformer block matrices, with positional embeddings excluded from the layerwise trap count.

Analysis choice	**Convention**
Diagnostic	Shuffled-spectrum outlier detection with WeightWatcher
Included modules	Transformer block matrices
Excluded module	Positional embedding layer `wpe`
Reason for exclusion	`wpe` is an embedding table, not a standard learned linear map
Layerwise output	Trap counts per analyzed transformer matrix

C Additional MLP Probes and Interventions

The randomized trap is a diagnostic of an exceptional spectral sector after entry permutation. To intervene on the model, we retain the permutation, edit the separated shuffled singular mode, invert the permutation, and reconstruct the original layer coordinates. The visualizations below are descriptive probes of the corresponding MLP checkpoint; they are not used to infer a unique mechanism from the spectrum alone.

C.1 Structure-destroying probes expose brittle output dependence

Figure 6 compares the anti-grokking MNIST–MLP on original MNIST, independently pixel-shuffled MNIST, and matched Gaussian probes. On original MNIST, the confusion matrix remains structured. Once spatial digit structure is destroyed, predictions concentrate on a small number of classes, especially class 5. This shows that the checkpoint is brittle to probes that retain only low-order marginal information.

The matched Gaussian inputs contain no class information or coherent strokes, yet the model again predicts almost exclusively class 5, with nearly constant confidence. The supported conclusion is that a low-information probe ensemble can activate a highly concentrated output response. The figure alone does not establish that input norm is the unique causal variable, nor does it show that the displayed leading vector is itself the complete anti-grokking mechanism.

The shuffled trap is not the learned pattern itself: entry permutation has destroyed row/column organization. Mapping an intervention back through the saved permutation identifies a sector of the original layer whose modification can be tested behaviorally. Any stronger claim about the feature represented by that sector requires the intervention, not the randomized spectrum alone.

C.2 Paired validation against in-distribution MNIST probes

We evaluate the exact final `WD=0` anti-grokking checkpoint, whose baseline test accuracy is 50.33%. For each of its 36 detected traps, we hold the one-trap-edited model fixed and compute JSD on both (i) 8,192 independent Gaussian inputs matched to the empirical MNIST input mean and variance and (ii) all 10,000 official MNIST test images. Thus the only change within each pair is the probe distribution, rather than the detected mode or edit realization. We use $T = 1$ and treat Spearman correlation as primary because absolute JSD scales can differ between probe distributions.

Table 12: **Paired Gaussian-JSD versus MNIST-JSD validation.** Correlations compare JSD scores for the same one-trap edits across all detected layers at the final `WD=0` MLP checkpoint.

Trap group	n	**Spearman** ρ (p)	**Pearson** r (p)
All detected layers	36	0.825 $(< 10^{-4})$	0.859 $(< 10^{-4})$

These results support random probes as a task-agnostic screen for functional activity. They do not make JSD sign-aware: labels and direct task evaluation remain necessary to determine whether removing a mode helps or harms generalization.

C.3 Intervening on the trap direction changes behavior

Figure 8 shows that the detected FC1 sector is not merely a visual curiosity. Replacing the mapped direction with a matched Gaussian random vector changes the test-set confusion structure and reduces the strong class-5 bias. This establishes that the edited sector is behaviorally active, although it does not by itself identify the full mechanism of anti-grokking.

C.4 Continued-training intervention

We intervene during a further 3,000 epochs of optimization. All conditions start from the same anti-grokking checkpoint with train/test accuracy $99.2\%/50.33\%$ and continue with AdamW. We compare ordinary continued training, the correction-budget-matched largest-weight edit from Section 4.5, and targeted removal of detected traps during training.

Table 13: **Behavior after 3,000 further epochs.** All conditions start at train/test accuracy $99.2\%/50.33\%$. Final outliers are mean $\pm$ SD over repeated entry permutations.

Condition	Best test	Final test	Final train	Final outliers
AdamW, no edit	50.79%	50.75%	99.2%	37.8 ± 2.6
Largest-$\lvert w\rvert$, same K, matched $\lVert \Delta W \rVert_F$	58.99%	58.92%	98.0%	10.6 ± 0.5
Detected traps removed during training	$\mathbf{65.29\%}$	$\mathbf{64.46\%}$	$\mathbf{100.0\%}$	$\mathbf{2.2 \pm 0.8}$

The scalar control is not null: it improves the optimization trajectory relative to no edit. Targeted removal nevertheless adds 6.30 percentage points in best test accuracy and 5.54 points in final test accuracy beyond that matched edit, restores 100% training accuracy, and leaves substantially fewer residual outliers. Sustained targeted removal therefore changes both the residual spectrum and the optimization trajectory. It does not imply that every individual trap is harmful; the single-mode ablations show heterogeneous effects.

C.5 Leading Eigenvectors, Receptive Fields, and Structural Outliers

For a weight matrix $\mathbf{W} \in \mathbb{R}^{N \times M}$, write

$$\mathbf{X}(\mathbf{W}) := \frac{1}{N}\mathbf{W}^\top \mathbf{W}, \qquad \lambda_{\max}(\mathbf{W}) := \lambda_{\max}\big(\mathbf{X}(\mathbf{W})\big). \tag{31}$$

If $\mathbf{W} = U\mathbf{\Sigma}\mathbf{V}^\top$ is the thin SVD, then the eigenpairs of $\mathbf{X}(\mathbf{W})$ are $(\sigma_k^2/N, \mathbf{v}_k)$, so the leading right singular vector identifies the dominant direction in the input space of the layer. In the MNIST–MLP, this makes the first-layer leading vector directly visualizable in pixel coordinates.

Figure 10 shows that anti-grokking introduces visibly extreme coordinates into the weight distributions of all three layers. This motivates, but does not replace, the explicit largest-weight controls in Table 3.

Figure 9 shows a complementary descriptive change: rows of $\mathbf{W}_1$ selected by leading-vector localization evolve from noise-like patterns to digit-like receptive fields in anti-grokking. These same rows have large projection coefficients when the mapped direction is expanded in the dominant eigenvectors of $\mathbf{X}(\mathbf{W}_1)$. Localization alone has limited and experiment-dependent predictive value for harmfulness, so task relevance is established by JSD and ablation rather than by visual form.

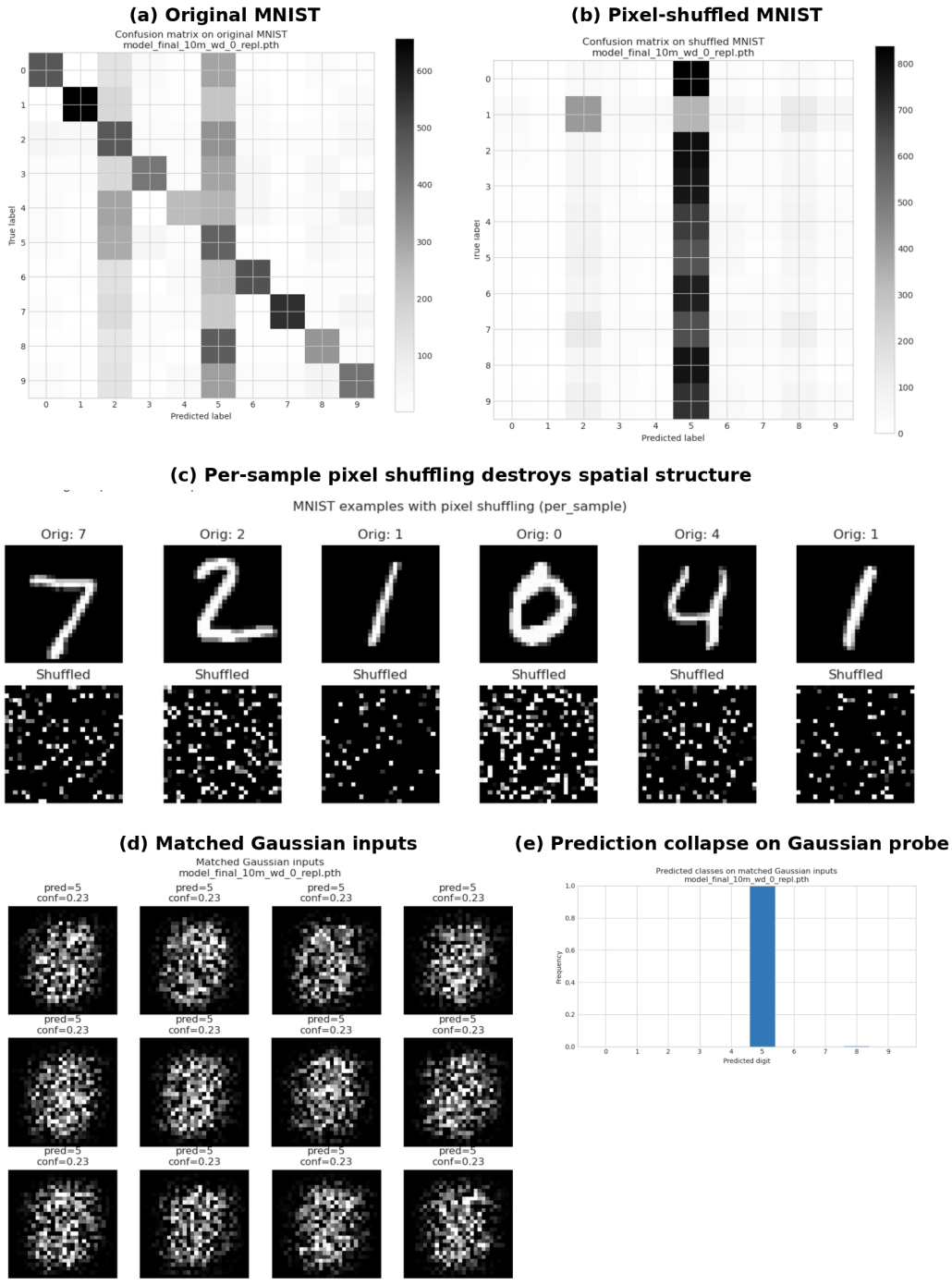
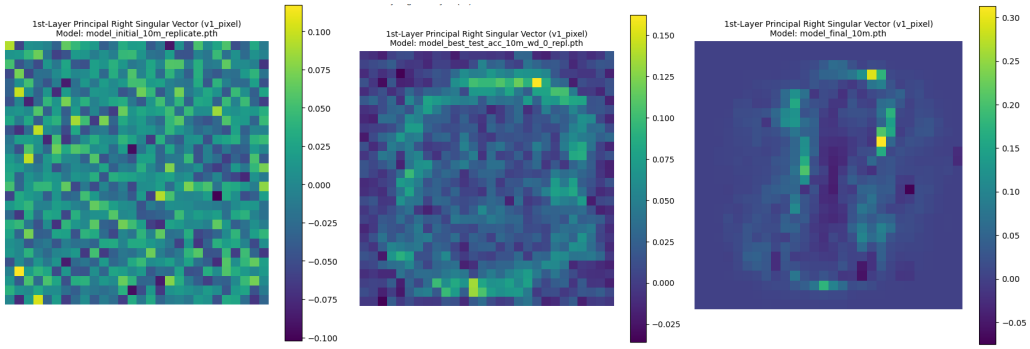


Figure 6: Structure-destroying probes for the anti-grokking MNIST–MLP. (a) The original MNIST confusion matrix remains partially structured. (b–c) After per-sample pixel shuffling, spatial digit information is destroyed, but predictions collapse toward a small set of labels. (d–e) Matched Gaussian probes contain no digit structure, yet the model predicts almost entirely class 5 with nearly constant confidence. Together, these probes show brittle output concentration after spatial structure is removed; they do not identify a unique causal statistic.



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

Figure 7: **Leading FC1 directions across phases.** Largest right singular vector $\mathbf{v}^{(1)}$ of $\mathbf{W}_1$ in pixel space. The direction evolves from an unstructured pattern in pre-grokking, to a smoother template during grokking, to a localized digit-like pattern in anti-grokking. This visualization is descriptive; causal relevance is evaluated by the matched intervention in Fig. 8.

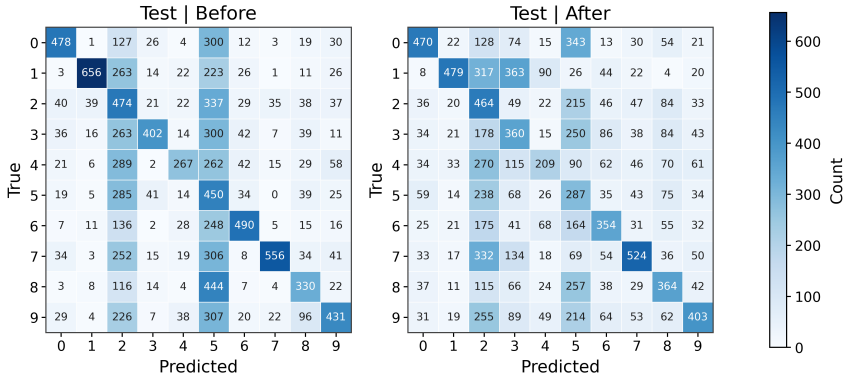
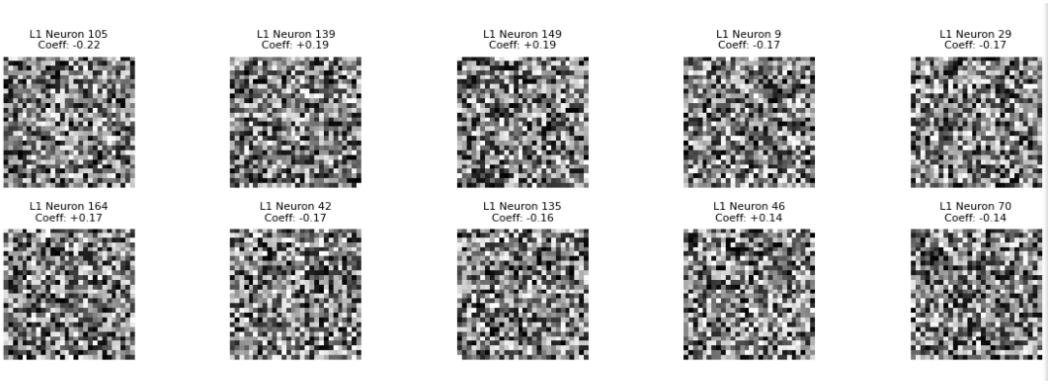
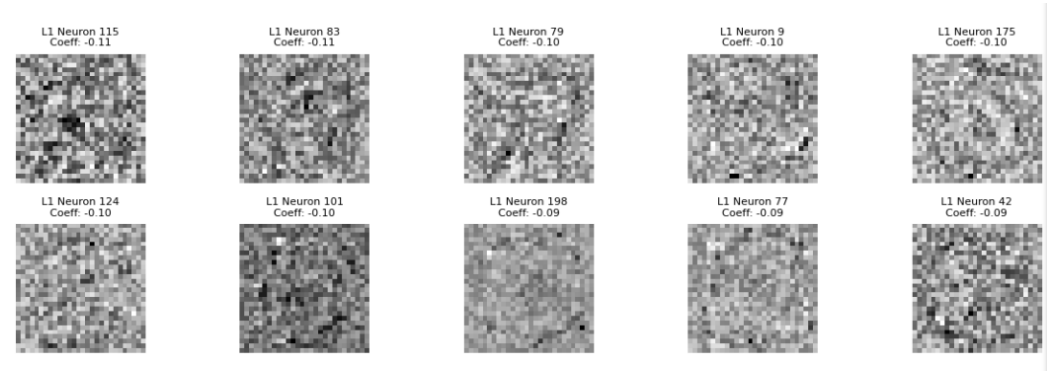


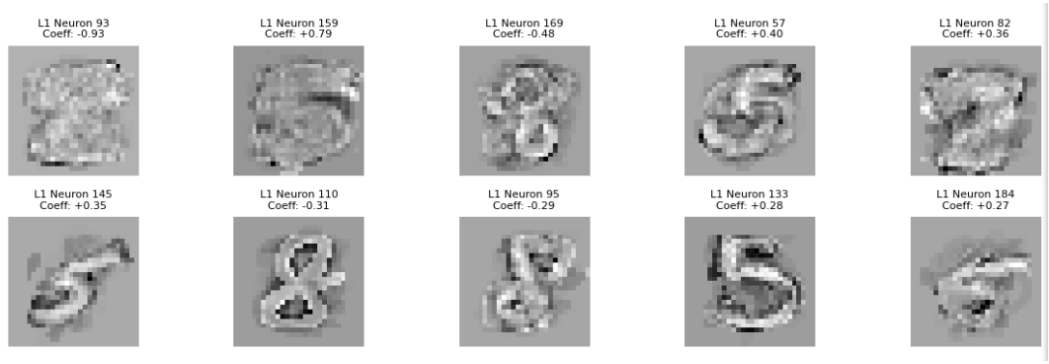
Figure 8: **Trap intervention in the anti-grokking phase.** (a) Confusion matrix for the full model in anti-grokking. (b) Confusion matrix after replacing the FC1 trap direction with a matched Gaussian random vector. The strong bias toward one prototype class (i.e., 5) is substantially reduced.



(a) Top-10 rows of $\mathbf{W}_1$ (pre-grokking)



(b) Top-10 rows of $\mathbf{W}_1$ (grokking)



(c) Top-10 rows of $\mathbf{W}_1$ (anti-grokking)

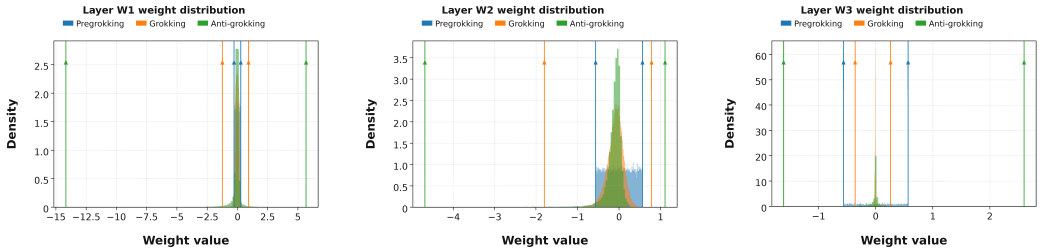
Figure 9: **Rows of $\mathbf{W}_1$ selected by leading-vector localization.** The selected receptive fields are noise-like in pre-grokking, remain diffuse around peak generalization, and become recognizable digit templates in anti-grokking.

C.6 Is the distribution of weights in anti-grokking simply very heavy-tailed?

If anti-grokking is accompanied by atypically large entries, one might inspect only the largest weights. That is not equivalent to the spectral edit: at the analyzed checkpoint, changing the same K largest coordinates with the same Frobenius correction budget leaves 37.6 ± 2.4 outliers compared with 38.2 ± 2.4 without editing, and hard-zeroing them still leaves 22.8 ± 3.6 (Table 3). Direct tail fitting poses a separate problem. The histograms may be Laplace-like, weakly heavy-tailed, or finite mixtures; these alternatives cannot be distinguished by visual inspection alone and require formal fitting.

Instead we use the fitted MP law as a nonparametric spectral baseline for the entry-wise randomized layer. The question is: after learned row/column arrangement is destroyed, how much randomized spectral mass remains beyond the fitted bulk and finite-size edge? A good fit is empirical evidence that this spectral observable follows the baseline; a separated edge mode is evidence that it does not. Neither observation, by itself, settles the asymptotic scalar tail law or the task effect.

The diagnostic does not require an infinite-variance law or a power-law exponent below two. Laplace distributions have finite moments and are compatible with the standard finite-moment regime; Gaussian-like and Laplace-like tensors are both observed in trained CNNs (Luqman et al., 2025). The empirical controls further show that the detected signal is not exhausted by one rank-one mean contribution or by the identity of the largest K weights.



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

Figure 10: **Weight distributions with extreme coordinates across checkpoints.** Anti-grokking introduces structural outliers into the weight values of all three layers. These extreme coordinates are consistent with the shuffled-spectrum outliers counted as Correlation Traps.

D MLP Perturbation Study

Do Traps Simply Reflect the Initialization Scale? As an additional robustness check, we repeat the MLP analysis across alternative initialization multipliers α_{init}. Changing α_{init} alters the overall weight scale and phase timing, but trap counts remain near zero in pre-grokking and grokking and become nonzero only in anti-grokking.

Protocol and phase definitions. For this study we use phase boundaries pre-grok (10^2–5×10^4 steps), grok (5×10^4–5×10^5 steps), and anti-grok ($> 5 \times 10^5$ steps). The reported statistics are mean±std over sampled checkpoints within each phase.

Table 14: Accuracy by phase under perturbation.

Phase	Train Accuracy (mean±std)	Test Accuracy (mean±std)
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 15: Global ℓ_2 norm by phase under perturbation.

Phase	WeightNorm (mean±std)
Pre-grok	17.52 ± 0.00
Grok	16.80 ± 0.22
Anti-grok	27.64 ± 8.39

Table 16: 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	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{4.56 \pm 3.03}$

The same pattern reappears: trap counts remain zero during pre-grokking and grokking and rise only in anti-grokking, whereas the global ℓ_2 norm changes throughout and does not by itself distinguish the phases.

E SELF-AVERAGING, CONCENTRATION, AND WHY MP IS THE RIGHT NULL

E.1 STATISTICAL-MECHANICS LANGUAGE VERSUS STATISTICAL CONCENTRATION

Self-averaging is always a property of an *observable*, not of a matrix in the abstract. In statistical mechanics one studies sample-to-sample fluctuations of quantities such as free-energy density, magnetization, or overlap. In statistics and learning theory the same idea appears as concentration: an observable A_N self-averages if $A_N - \mathbb{E}A_N \to 0$ in probability, with $\mathrm{Var}(A_N) \to 0$ providing a sufficient L^2 criterion. Non-self-averaging means that order-one fluctuations persist for the specified observable as system size grows.

Historical statistical-mechanics view. The spin-glass interpretation of neural-network overfitting is older than modern concentration language. The Hopfield and Amit–Gutfreund–Sompolinsky models already showed in the 1980s that associative memories develop a glassy phase with many metastable, sample-specific states once interference from stored patterns becomes strong enough Hopfield (1982); Amit et al. (1985). The Gardner program then turned this into a geometric theory of learning constraints: one studies the volume of interactions compatible with many patterns, and the overconstrained regime is naturally described by a glassy landscape rather than a single clean retrieval state Gardner (1988). By the 1990s, statistical-mechanics analyses of learning and regularization were already interpreting poor generalization and overfitting as movement into such glassy, sample-dependent states, with weight decay or early stopping acting as ways of keeping the system away from that regime Seung et al. (1992); Bös (1998).

Modern concentration-theoretic view. We use the same qualitative picture in observable-specific probabilistic language. The trained matrix is one realization of quenched disorder induced by the training sample and optimization path. When the randomized entry distribution belongs to a finite-moment MP/TW class, its spectral bulk and ordinary edge fluctuations should concentrate around a typical large-layer baseline. An exceptional element distribution, edge, or task-relevant sector can converge slowly or fail to concentrate. The paper tests this possibility empirically; it does not claim a full disorder-ensemble theorem for trained neural networks.

This observable-specific viewpoint matters here. A shuffled weight matrix can have well-behaved bulk averages and an exceptional edge observable. We do not say that an entire layer is non-self-averaging in every statistic, nor that a trap automatically makes generalization error fail to concentrate. We say that traps nominate exceptional sectors that can matter when they overlap a task-relevant direction; output-space JSD and direct intervention test that overlap empirically.

E.2 WHY MP IS THE APPROPRIATE RANDOMIZED BASELINE

For covariance-type random matrices with independent, well-behaved entries, the empirical spectral density concentrates around the Marchenko–Pastur law. The null also carries a vector statement: bulk singular/eigenvectors are delocalized rather than concentrated on a tiny coordinate set. Our permuted entries are exchangeable without replacement, not independent, so MP/TW is used conditionally and checked by fit quality rather than asserted as an exact data-generating model.

The relevant right-edge benchmark is not the fitted bulk edge alone but the bulk edge together with Tracy–Widom-scale edge fluctuations. The Baik–Ben Arous–Péché transition then provides the conceptual interpretation: once an eigenvalue cleanly separates past that edge scale, it should no longer be viewed as an ordinary random fluctuation but as a structural outlier. That is exactly how Correlation Traps are used in the main paper.

E.3 WHY MODERATE TAILS AND MULTI-ENTRY STRUCTURE ARE ENOUGH

The diagnostic is implemented as a uniform entry-wise permutation of the learned multiset. The dependence induced by sampling without replacement is precisely why we report the phase-level KS distances and repeat the analysis over fresh permutations. The scientific conclusion is empirical: the fitted null is adequate in the trap-free phases, fails sharply at collapse, and produces a stable phase classification under re-randomization.

Likewise, the diagnostic does not require extremely heavy tails or a sub-Gaussian element law. A Laplace distribution is heavier-tailed than a Gaussian but has finite moments. A single large entry can create an outlier, but the matched largest-weight controls show that the observed anti-grokking signal is not explained by the same K extreme coordinates alone. The empirical histograms may be Laplace-like, weakly heavy-tailed, or finite mixtures; this paper does not infer an asymptotic tail exponent from them.

F Robustness

This appendix records the compact robustness summaries that support two paper-level claims: the three-phase anti-grokking trajectory is not a single-run artifact, and the shuffled-spectrum diagnostic is not driven by one lucky permutation. These checks use the same benchmarks and checkpoints already analyzed in the main paper. Their role is technical rather than conceptual: they confirm that the story is unchanged under modest seed-to-seed and shuffle-to-shuffle variation.

F.1 Sampling with versus without replacement

Our primary randomization uniformly permutes each learned layer's entries without replacement. To test whether this choice drives the result, we compare it with drawing the same number of entries with replacement from the layer's learned scalar-weight multiset. The paired comparison uses five MLP checkpoints from each phase, 25 matched random seeds, all three analyzed weight matrices, and the identical `WeightWatcher` MP-density fit and MP/TW cutoff, for a total of 2,250 layer-level spectral fits.

Across the 15 checkpoint means, the two randomization schemes agree almost perfectly:

$$r_{\text{Pearson}} = 0.9993, \qquad \rho_{\text{Spearman}} = 0.9519. \tag{32}$$

Table 17 shows that the phase-level trap signal is essentially unchanged. Sampling with replacement slightly increases variability but does not systematically raise or lower the trap count.

Table 17: **Trap counts under sampling with and without replacement.** Values are phase-level means $\pm$ SD under the paired protocol. The final row gives paired checkpoint-bootstrap 95% intervals for the difference between schemes.

Randomization	Pre-grok	Grok	Anti-grok
Without replacement	0.00 ± 0.00	2.33 ± 0.64	33.99 ± 2.89
With replacement	0.00 ± 0.00	2.28 ± 0.55	33.76 ± 3.57
With minus without	0.00	-0.05	-0.23
Paired bootstrap 95% interval	$[0, 0]$	$[-0.18, 0.08]$	$[-1.21, 0.62]$

Sampling with replacement also perturbs the empirical marginal distribution in each realization: during anti-grokking, the layer standard deviation changes by 2.67% on average and the maximum absolute weight changes by 3.37% on average. In contrast, permutation without replacement preserves the mean, variance, quantiles, and extrema to numerical precision. We therefore retain without-replacement permutation as the cleaner conditional reference, while the paired control shows that the phase separation and checkpoint-level trap counts do not materially depend on this choice.

F.2 Phase stability across random seeds

Table 18 summarizes qualitative robustness across the analyzed seed runs. Collapse time changes across seeds, but the ordering is unchanged: pre-grokking remains trap-free, grokking remains trap-free or nearly so, and anti-grokking begins only after successful generalization and coincides with clearly nonzero trap counts.

F.3 Stability across repeated entry-wise randomizations

Because the diagnostic permutes each layer entry-wise, exact trap counts vary across randomizations. In the trap-free phases, repeated shuffles remain at or near zero and do not create phase ambiguity.

Table 18: **Qualitative robustness across random seeds.** The onset step of anti-grokking varies across seeds, but the three-phase structure and trapped-versus-untrapped separation remain unchanged.

Phase	Observed behavior across seeds	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 or no traps	Stable generalization regime
Anti-grokking	Test accuracy declines under continued optimization while trap counts become clearly nonzero	Stable trapped overfitting regime

Table 19 reports the quantitative anti-grokking summaries available from the reviewer-requested controls and continued-training intervention.

Table 19: **Repeated-randomization summaries.** Means and standard deviations are over fresh entry permutations at the indicated anti-grokking checkpoints.

Checkpoint / condition	Fresh shuffles	Total outliers, mean $\pm$ SD
One-shot, unmodified	50	38.2 ± 2.4
One-shot, layer-wise mean subtraction	50	36.4 ± 1.8
After 3,000 epochs, no intervention	repeated	37.8 ± 2.6
After 3,000 epochs, trap removal	repeated	2.2 ± 0.8

These summaries support phase classification rather than exact count invariance. Pre-grokking and grokking remain at or near zero, anti-grokking remains clearly nonzero, and targeted removal leaves far fewer residual modes than continued training without intervention.

G LIMITATIONS

Our controlled evidence comes from three long-horizon grokking settings: an MNIST MNIST–MLP, modular addition (MODADD), and a GPT2–COMPOSE-style training trajectory. These experiments expose the pre-grokking, grokking, and anti-grokking phases clearly, but they do not exhaust the space of architectures, optimizers, datasets, losses, or seeds. We do not claim that every kind of overfitting produces traps.

The randomized null is approximate. Entry permutation gives exchangeable sampling without replacement, not independent entries. MP/TW is appropriate when the observed multiset has controlled mean, finite variance, and no marginal domination. The phase-level KS check, centering control, and repeated shuffles test adequacy empirically but do not turn the finite-population ensemble into an i.i.d. model.

JSD is label-free but not sign-aware. It ranks the output influence of a mode on a chosen probe distribution; a poor probe can miss task-relevant behavior, and only labeled intervention can establish whether removal helps or hurts. The GPT-OSS results are screening evidence only and do not establish harmful behavior. Trap counts should not be used as universal model-quality, overfitting, or safety scores without the same probe-based and task-based calibration used in the controlled experiments.

H REPRODUCIBILITY ARTIFACTS

The submitted manuscript archive contains the LaTeX source and figure assets. We do not claim that this archive alone is an end-to-end executable release. The anonymized reproducibility artifact

accompanying the final submission will include: (i) training scripts for the MNIST–MLP and modular-addition benchmarks; (ii) executable trap detection, removal, KS, and figure-generation scripts; (iii) the `WeightWatcher` v0.7.5.5 version lock; (iv) layer-selection rules, seeds, fit and threshold parameters; (v) representative checkpoints or regeneration instructions; and (vi) example commands and environment details. The datasets and core analysis tool are public: MNIST is used for the MNIST–MLP benchmark, and trap analysis uses `WeightWatcher` Martin (2018-2024).

I ASSETS, LICENSES, AND ATTRIBUTION

The paper uses only existing public assets. The MNIST–MLP benchmark uses MNIST, whose commonly distributed dataset license is Creative Commons Attribution-Share Alike 3.0. The shuffled-spectrum analysis uses the open-source `WeightWatcher` package, version v0.7.5.5, distributed under the Apache License 2.0. The exploratory frontier-model analysis uses the public `gpt-oss-20b` and `gpt-oss-120b` open-weight releases, which are made available under the Apache License 2.0 and the corresponding OpenAI usage policy. We do not introduce a new dataset.