THE GROKKED ILLUSION: TRUE EQUILIBRIUM MITIGATES CATASTROPHIC FORGETTING

Anonymous authors
Paper under double-blind review

ABSTRACT

While neural networks are typically evaluated by their training and test performance, these metrics do not reveal how robust a learned representation is. Recent studies have shown that solutions occupying larger volumes in parameter space, as quantified by Boltzmann entropy, often exhibit superior generalizability compared to those reached by conventional optimization, a phenomenon known as the *high entropy advantage*. Here we ask whether this advantage persists beyond generalization. Specifically, we investigate models' robustness, the ability to retain the learned knowledge when the model is subsequently trained to acquire new information. Using grokking in modular arithmetic as a controlled setting, we design a noise injection experiment to evaluate the robustness difference between AdamW-trained transformers and high-entropy model sampled from Wang-Landau Molecular Dynamics with identical saturated performance. By forcing both models to fully remember new data with random labels, we find that AdamW-trained models suffer from catastrophic forgetting, with original task test accuracy dropping from 100% to about 77%, whereas the high-entropy models maintain approximately 95% test accuracy. We term this hidden fragility behind apparent generalization the "grokked illusion." Through singular value decomposition of the neural network weights, we discover that high-entropy neural networks possess significantly higher effective rank in attention and MLP layers both before and after noise injection, which is consistent with the concept that richer feature representations act as a buffer against catastrophic forgetting. Our findings demonstrate that perfect generalization does not imply equal robustness, offering a new perspective on what makes a trained model robust to interference.

1 INTRODUCTION

Generalization performance has long been a key metric for evaluating machine learning models, reflecting the ability of neural networks (NNs) to learn knowledge and apply it to unknown data. Yet a model that generalizes perfectly to a test set may still be fragile: parameter updates caused by memorizing new input data may lead to a catastrophic collapse of its previously acquired knowledge. This disconnect between generalization and robustness has been observed across a wide range of settings, from adversarial examples (Goodfellow et al., 2015) to catastrophic forgetting in continual learning (McCloskey & Cohen, 1989), and it raises a fundamental question: what properties of a NN determine not only its generalizability, but also its robustness to interference?

Modern NNs often have a large number of parameters, enabling them to perfectly fit finite training data and even random noise (Zhang et al., 2017). One mainstream research direction attempts to characterize NN states through the geometric structure of the parameter space. A classic conjecture in this direction is that flatter minima in the loss landscape correspond to higher generalizability (Hochreiter & Schmidhuber, 1997). However, due to the high dimensionality of the landscape, local flatness metrics such as the maximum eigenvalue of the Hessian matrix are usually difficult to calculate and cannot express the full picture of the loss landscape(Ghorbani et al., 2019; Dinh et al., 2017). Even worse, the sharpness quantities computed from the Hessian can be arbitrarily reshaped by simply rescaling of the weights, without impacting NNs' generalization performance (Dinh et al., 2017). A recent study introduced the Boltzmann entropy of a NN state as a global measure of the volume in parameter space corresponding to a given training loss and test performance (Yang et al., 2026b). Their work reveals the high-entropy advantage: high entropy states that occupy

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 larger volume in the parameter space exhibit superior generalizability compared to conventionally trained solutions. This observation leads naturally to a hypothesis about robustness: if a high-entropy solution resides within a larger basin in parameter space, then when its parameters are displaced by external perturbations, the resulting state is more likely to remain within the same high-entropy region, thereby preserving its learned functionality. In other words, the same property that confers a generalization advantage may also create a robustness advantage.

To test this hypothesis while controlling for the confounding effect of generalization performance itself, we choose modular arithmetic (Power et al., 2022) as our experimental setting. In this task, both conventionally trained Transformer models and equilibrium states (states with the highest entropy under a given training loss) can achieve identical perfect test accuracy (Zhang et al., 2026), allowing us to isolate the effect of entropy on robustness. We design a noise injection experiment in which a NN that has already achieved perfect generalization in the original task must additionally memorize new noisy data. Despite the continued availability of the original training data during noise injection, AdamW-trained NNs exhibit a severe drop in original task test accuracy from 100% to about 77% in the most challenging condition. In stark contrast, the equilibrium NNs sampled via Wang-Landau Molecular Dynamics (WLMD) maintain approximately 95% test accuracy on the original task after memorizing the same noisy data. We term this phenomenon the "grokked illusion", where solutions appear to grok (saturated test performance) for the same task can present different levels of robustness. Through singular value decomposition (SVD) of the NN weights, we find that equilibrium NNs possess significantly higher effective rank in attention and MLP layers both before and after noise injection, which is consistent with the concept that richer feature representations can act as a buffer against catastrophic forgetting.

Our main contributions are three-fold:

1. We identify the "grokked illusion," a phenomenon in which NNs that achieve perfect generalization via conventional training nevertheless suffer from catastrophic forgetting when faced with new noisy inputs.
2. We demonstrate the high-entropy robustness advantage: high-entropy states maintain significantly stronger retention of prior knowledge than conventionally trained NNs.
3. We show through singular value analysis that high-entropy states exhibit systematically higher effective rank, providing a structural correlate of their stronger robustness during continual learning.

2 METHODOLOGY

2.1 TRAINING PROTOCOLS

We adopt the modular arithmetic task $(x + y \bmod 67)$ as our primary benchmark, following the setup of (Nanda et al., 2023; Zhang et al., 2026). The model is a single-layer Transformer with vocabulary size $d_{\text{vocab}} = 68$, embedding dimension $d_{\text{model}} = 128$, MLP hidden dimension $d_{\text{mlp}} = 512$, $d_{\text{head}} = 32$, and 4 attention heads, amounting to approximately 2.15×10^5 parameters, which is severely overparameterized relative to the training set size of 2244 samples (approximately half of the 67×67 possible input pairs).

To obtain high-entropy NNs state, we employ the Wang-Landau Molecular Dynamics (WLMD) method (Wang & Landau, 2001) introduced in prior work (Yang et al., 2026b; Zhang et al., 2026). The WLMD algorithm samples the Boltzmann entropy landscape $S(\mathcal{L}_{\text{train}}, \mathcal{A}_{\text{test}}) = \log V(\mathcal{L}_{\text{train}}, \mathcal{A}_{\text{test}})$, where V denotes the volume in the parameter space corresponding to a given training loss $\mathcal{L}_{\text{train}}$ and test accuracy $\mathcal{A}_{\text{test}}$. We restrict the sampling to configurations with weight norm $\|w\| = 30$, following the "Goldilocks zone" identified in (Liu et al., 2022), and used the "entropy bias" method (Zhang et al., 2026) to constrain the sampling to the range $\ln(\mathcal{L}_{\text{train}}) \in [-20, 1]$, which ensure that the entropy landscape within the low training loss range of interest converges correctly. The simulation is run for 1.2×10^8 WLMD epochs until the histogram flatness criterion is satisfied, indicating convergence of the entropy landscape. From the converged equilibrium curve, we select a NN state that achieves test accuracy $= 100\%$ with $\ln(\mathcal{L}_{\text{train}}) \approx -1$. We refer to such states as *equilibrium states* (or high-entropy states), i.e., configurations that maximize entropy for a given training loss.

For comparison, we train NNs using the AdamW optimizer with the same weight norm constraint $\|w\| = 30$. The weight norm is enforced after each parameter update via linear scaling following (Liu et al., 2023): after each AdamW update, we rescale all parameters by a factor such that the resulting NN satisfies $\|w\| = 30$. We use a learning rate of 1×10^{-2}, weight decay of 1.0, and full-batch training for 120 epochs, producing a NN with $\ln(\mathcal{L}_{\text{train}}) \approx -2$ and test accuracy 100%. Throughout this paper, we denote these two types of NNs as *WLMD-equilibrium* (WLMD) and *AdamW-goldilocks* (AdamW), respectively. Both are trained under identical weight norm constraints and achieve perfect test accuracy, allowing us to isolate the effect of entropy on robustness while controlling for generalization performance and weight scale.

We also examined the robustness of AdamW trained on the full dataset (including test samples) under the same weight norm constraint ($\|w\| = 30$) and standard AdamW-trained NNs, with results presented in the Appendix. Full details of the computing infrastructure, software dependencies, and hyperparameter configurations are provided in the Appendix.

2.2 Noise Injection Protocol

In this work, we define *robustness* as the ability of a NN to retain previously learned knowledge when its parameters are subsequently updated to accommodate new information. We quantify robustness by measuring the NN's test accuracy on the original task after it has fully memorized the injected noisy data (i.e., achieving $> 99.8\%$ training accuracy on the combined dataset). To evaluate this property while controlling for the confounding effect of generalization performance, we design a noise injection experiment in which a NN that has already mastered the original modular arithmetic task must additionally memorize new noisy data. We first expand the vocabulary dimension of the trained NN from 68 to 136 by duplicating the embedding and output projection layers, and randomly initializing the newly added parameters. The original input tokens occupy indices $0, \ldots, 67$, while the inputs of the noise tasks occupy the shifted token range $68, \ldots, 104$ (the 37 numerical symbols of the mod-37 tasks); the corresponding function symbol is placed at index 135. The remaining indices $105, \ldots, 134$ are left unused as redundant vocabulary dimensions. This expansion preserves the NN's existing functionality on the original task while equipping it with the capacity to process new inputs.

We refer to the newly introduced data as *noise* in the sense that its input–label relation differs from the original task, even when it follows a regular rule of its own. We construct three types of noisy datasets, each containing 500 training samples drawn from shifted token range, with labels read out from original number tokens ($0, \ldots, 66$ for random labels and $0, \ldots, 36$ for the two structured rules):

- **Random noise**: inputs are randomly generated;
- $\mathbf{x^2 + y} \bmod \mathbf{37}$: inputs follow a structured arithmetic rule with modulus 37, offering partial structural similarity to the original task;
- $\mathbf{x + y} \bmod \mathbf{37}$: inputs follow the same additive rule as the original task but with a different modulus, representing the closest structural relationship among the three noise types.

These three noise types form a spectrum of structural similarity to the original task, ranging from completely unrelated to progressively more aligned, which allows us to examine how the nature of new knowledge affects the retention of previously learned representations.

The expanded NN is then fine-tuned on a mixed training set comprising the original 2244 training samples and the 500 noisy samples. We use the AdamW optimizer with a learning rate of 1×10^{-4}, full-batch updates, and a maximum of 5000 epochs. Training is terminated when the model achieves $> 99.8\%$ training accuracy on the combined dataset, indicating that the NN has fully memorized both the original and the noisy data. In practice, this criterion is reached within 4500 epochs across all settings. We monitor the NN's test accuracy on the original task throughout this fine-tuning process. Importantly, the training accuracy on the original task data remains at 100% during fine-tuning, confirming that the NN continues to fit the original training data. Nevertheless, as we will show, AdamW-trained NNs still exhibit a significant drop in original-task test accuracy, revealing that maintaining perfect fit to the training data does not guarantee the retention of generalizable knowledge. Throughout fine-tuning, all 2244 original training examples remain in every update, so the model continues to fit the original task exactly; we use "catastrophic forgetting" to denote the degradation of *held-out* accuracy on the original task under this rehearsal protocol.

2.3 Singular Value Decomposition and Effective Rank Analysis

To examine the spectral structure associated with the observed robustness differences, we perform singular value decomposition (SVD) on the weight matrices of individual layers before and after noise injection. For a weight matrix $W \in \mathbb{R}^{m\times n}$, let its singular values be $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > 0$, where $k = \min(m, n)$. We define the normalized singular value distribution as $\tilde{\sigma}_i = \sigma_i / \sum_{j=1}^{k} \sigma_j$. Following (Roy & Vetterli, 2007), the *effective rank* (ER) is defined as:

$$\mathrm{ER}(W) = \exp\left(-\sum_{i=1}^{k} \tilde{\sigma}_i \log \tilde{\sigma}_i\right),$$

which quantifies the number of significant singular directions and serves as a measure of the representational richness of the layer. A higher ER indicates a more uniform singular value spectrum, implying that the layer utilizes a broader set of feature dimensions.

We compute the effective rank for all weight matrices in the NN both before and after noise injection. Additionally, we visualize the singular value spectra by plotting the sorted singular value index against the corresponding singular values on a log-log scale, which we term the *Singular Value Spectrum* (SVS) plot. This visualization provides a direct comparison of the distributional shape of singular values between equilibrium and AdamW-goldilocks NNs.

To complement the effective rank analysis, we also compute the cosine similarity of the flattened parameter vectors between the NNs before and after noise injection, providing a global measure of parameter displacement. The cosine similarity is defined as:

$$\mathrm{Cos}(\theta_{\text{before}}, \theta_{\text{after}}) = \frac{\theta_{\text{before}} \cdot \theta_{\text{after}}}{\|\theta_{\text{before}}\| \|\theta_{\text{after}}\|},$$

where θ denotes the flattened concatenation of all trainable parameters. All SVD analyses are conducted on the weight matrices without bias terms, as biases contribute negligibly to the overall representational capacity in this architecture.

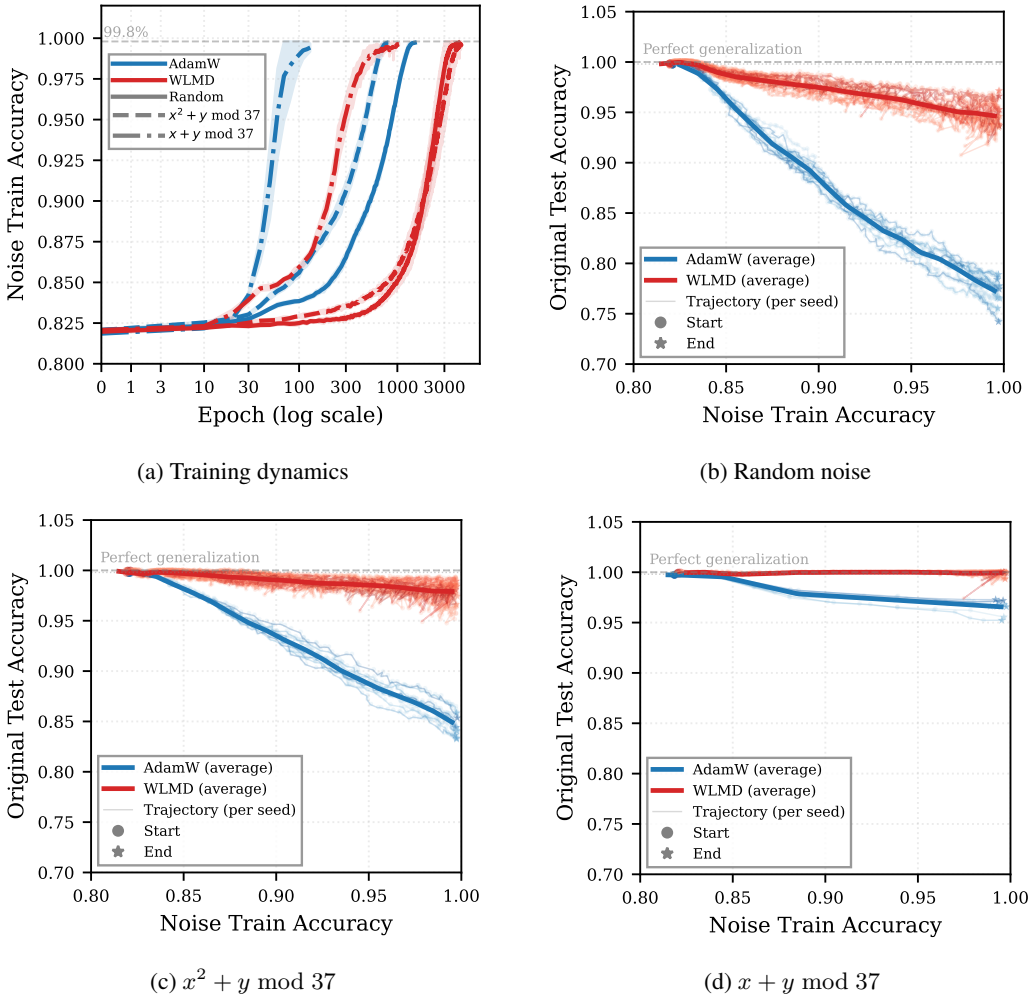
3 Experimental Results

We conduct noise injection experiments on four types of NN states: equilibrium state sampled by WLMD (denoted as WLMD-equilibrium NN), AdamW with constrained weight norm $\|w\| = 30$ (denoted as AdamW-goldilocks NN), standard AdamW without weight norm constraint (denoted as AdamW-standard NN), and AdamW trained on the full dataset (denoted as AdamW-full NN). Our main results focus on the comparison between the WLMD-equilibrium NN and the AdamW-goldilocks NN: both reach perfect test accuracy under the same weight norm constraint, which allows us to compare two states of identical test performance and weight scale. Results for the AdamW-standard and AdamW-full controls are provided in Appendix A.3.

3.1 Robustness Advantage of WLMD-Equilibrium Neural Networks

We evaluate the robustness of learned knowledge by measuring the original-task test accuracy during fine-tuning on mixed datasets containing both original training data and new noisy data. For each experimental condition, we run noise injection with ten different random seeds in the range [40,50).

Figure 1 presents the training dynamics for three types of noisy data: random noise, $x^2 + y \bmod 37$, and $x + y \bmod 37$. Figure 1a reveals that for all three noise types, both WLMD-equilibrium and AdamW-goldilocks NNs can achieve $> 99.8\%$ training accuracy on the mixed dataset, confirming that both types of NNs successfully memorize the injected noisy data. WLMD-equilibrium NNs consistently require more epochs to memorize the noisy data than AdamW-goldilocks NNs across all noise types, suggesting that NNs at the equilibrium are less sensitive to noise injection and more difficult to adapt. Those NNs' behavior on the original task diverges sharply. Figures 1b–1d show the (Noise Train Accuracy, Original Test Accuracy) trajectories for the three noise types. In the random noise condition (Fig. 1b), AdamW-goldilocks trajectories exhibit a severe drop from 100% to about 75% in original-task test accuracy. This phenomenon, which we term *grokked illusion*, reveals that for the NN generalized at tasks involving grokking through conventional training methods, fitting the



(a) Training dynamics

(b) Random noise

(c) $x^2 + y$ mod 37

(d) $x + y$ mod 37

Figure 1: Training dynamics and robustness trajectories. (a) Training accuracy on the combined dataset (original training data plus noise) versus epoch (log scale) for AdamW (blue) and WLMD (red) under all three noise types, with shaded areas indicating ± 1 standard deviation across 10 seeds. Solid/dashed/dash-dotted lines correspond to random, $x^2 + y \bmod 37$, and $x + y \bmod 37$, respectively. (b–d) Trajectories of (Noise Train Accuracy, Original Test Accuracy) during fine-tuning for random, $x^2 + y \bmod 37$, and $x + y \bmod 37$ noise, respectively, with mean trajectories highlighted in blue (AdamW) and red (WLMD). Gray dashed line at $y = 1.0$ marks perfect generalization on the original task.

training data alone does not guarantee the retention of its knowledge when new irrelevant information is introduced. In stark contrast, the WLMD-equilibrium NN maintains approximately 95% test accuracy on the original task after fully memorizing the random noise, demonstrating a substantial advantage in robustness.

For the structured noise types (Figs. 1c and 1d), the robustness gap persists though narrows as the noise becomes more structurally similar to the original task. Specifically, for $x^2 + y \bmod 37$, AdamW-goldilocks drops to approximately 84% test accuracy on the original task, while WLMD maintains above 98%. For $x + y \bmod 37$, where the underlying rule is identical in form to the original task, AdamW-goldilocks achieves approximately 97% and WLMD maintains about 100%. This monotonic trend indicates that the robustness advantage of WLMD-equilibrium NNs is consistently present across different types of new knowledge, and that the disruption to prior knowledge is inversely related to the structural proximity between the new and original tasks.

Layer	AdamW-goldilocks				WLMD-equilibrium			
	ER_{pre}	ER_{post}	ΔER	Cos	ER_{pre}	ER_{post}	ΔER	Cos
Embedding	63.8	55.3±0.4	-8.5±0.4	0.9962±0.0005	68.7	64.7±0.8	-4.0±0.8	0.9892±0.0018
Pos. Embedding	1.5	1.9±0.0	+0.4±0.0	0.9946±0.0011	2.2	2.2±0.0	-0.1±0.0	0.9997±0.0002
W_Q	2.2	2.4±0.2	+0.2±0.2	0.9964±0.0023	96.8	68.3±5.6	-28.5±5.6	0.8383±0.0379
W_K	7.2	6.9±0.1	-0.3±0.1	0.9857±0.0034	97.4	86.7±2.0	-10.7±2.0	0.9607±0.0109
W_V	33.6	35.1±0.4	+1.5±0.4	0.9778±0.0017	84.2	75.1±0.9	-9.1±0.9	0.9721±0.0033
W_O	36.2	29.3±0.1	-6.9±0.1	0.9978±0.0002	83.5	61.4±0.6	-22.1±0.6	0.9307±0.0039
W_{in}	48.1	41.4±0.3	-6.8±0.3	0.9960±0.0003	116.1	83.1±0.8	-33.0±0.8	0.8566±0.0073
W_{out}	33.7	44.6±0.5	+10.9±0.5	0.9454±0.0015	116.6	100.4±0.4	-16.2±0.4	0.9572±0.0015
Unembedding	70.7	62.0±0.6	-8.7±0.6	0.9451±0.0077	73.3	58.5±3.9	-14.7±3.9	0.9856±0.0082

Table 1: Effective rank (ER) before and after noise injection, change in ER (ΔER), and cosine similarity (Cos) for each layer. Post-injection quantities are reported as mean $\pm$ std over 10 seeds; pre-injection ER is from the single pre-trained checkpoint.

To rule out the possibility that this robustness advantage originates from the use of the test set in defining the NNs' entropy landscape, we additionally train a control network on the full dataset including the original test samples (AdamW-full). Under random noise this network retains only about 82% of the original-task accuracy, close to the 77% of the constrained AdamW state and far below the 95% of the equilibrium state. The same ordering holds for the two structured noise types. This suggests that even direct exposure to the test set during training cannot lead to a retention performance matching the equilibrium state. Detailed results are reported in Appendix A.3.

3.2 Singular Value Analysis and Spectral Interpretation

To examine how the two states differ spectrally, we perform singular value decomposition on all weight matrices of the AdamW-goldilocks and WLMD-equilibrium NNs before and after noise injection, with the random noise condition as the primary case study. The complete Singular Value Spectrum (SVS) plots for each noise and NN types are provided in the Appendix. Table 1 reports the effective rank (ER) before and after noise injection, the change in ER, and the cosine similarity (Cos) between the pre- and post-injection weight matrices for each layer across both NN types.

The effective rank analysis reveals a pattern that varies across layer types. For the four attention projection matrices (W_Q, W_K, W_V, W_O) and the two MLP layers ($W_{\text{in}}, W_{\text{out}}$), the WLMD-equilibrium NN exhibits significantly higher ER than the AdamW-goldilocks NN prior to noise injection. This indicates that the high-entropy state utilizes a substantially broader set of singular directions in these core layers, reflecting a richer feature representation of the learned task.

After noise injection, the effective rank decreases in most layers of both models as they adapt to accommodate the new noisy data. Notably, for the attention and MLP weight parameters, the absolute decrease in ER is larger for the WLMD-equilibrium NN than for the AdamW-goldilocks NN. Despite this greater reduction, the post-injection ER of the WLMD-equilibrium NN remains substantially higher than that of the AdamW-goldilocks NN across these layers. For the unembedding layer, the two NNs exhibit comparable ER prior to noise injection, but after injection, the WLMD-equilibrium NN shows a larger decrease, resulting in a post-injection ER slightly lower than that of AdamW-goldilocks. These observations suggest that high-entropy NNs possess an abundance of representational capacity in their attention and MLP layers, which serves as a buffer: even after sacrificing a larger number of effective feature dimensions to accommodate new information, they still retain sufficient richness to preserve the original task knowledge.

The cosine similarity between parameter vectors before and after noise-injection is higher for the AdamW-goldilocks NN compared to the WLMD-equilibrium NN, indicating that the WLMD-equilibrium NN undergoes substantially larger parameter displacement. This result aligns directly with the entropy perspective: because the WLMD-equilibrium NN corresponds to a high-entropy region of significantly larger volume in parameter space, such a state can accommodate greater parameter displacement while preserving the learned function, thereby retaining its generalizability on the original task. In contrast, the AdamW-goldilocks NN corresponds to a much narrower region, suffers from catastrophic forgetting when its parameters are displaced out of the low-entropy basin.

4 Related Work

4.1 Parameter-Space Geometry, Generalization, and Statistical Physics

The relationship between the geometric structure of the parameter space and a NN's generalization has been a central theme in deep learning theory, the classic conjecture being that flatter minima in the loss landscape correspond to better generalization (Hochreiter & Schmidhuber, 1997). However, subsequent work has revealed that sharp minima can also generalize well (Dinh et al., 2017), and that flatness measures based on the Hessian can be arbitrarily manipulated by rescaling the parameters without affecting generalization performance.

Statistical physics has provided a powerful lens for these phenomena: Chaudhari et al. introduced entropy-SGD, linking the physics energy landscape to generalization through a local entropy objective that biases optimization toward wide valleys (Chaudhari et al., 2019), and Bahri et al. reviewed statistical-physics approaches to deep learning (Bahri et al., 2020).

Within this tradition, the Boltzmann entropy of NN states, defined as the volume of parameter space corresponding to a given training loss and test performance, has been introduced recently as a global measure that complements local flatness metrics by capturing the structure of the solution space (Yang et al., 2026b). Yang et al. established the high-entropy advantage in generalizability: NN states occupying larger parameter space volumes exhibit superior generalization compared to conventionally trained solutions, validated across tasks including arithmetic reasoning, tabular data, image recognition, and language modeling (Yang et al., 2026b). Subsequent work by Zhang et al. examined this advantage for grokking on modular arithmetic tasks, framing grokking as a computational glass relaxation process and demonstrating that the high-entropy advantage is particularly significant in this setting (Zhang et al., 2026). The present study further extends this line of research by shifting the focus from generalization (measured by test metrics) to robustness (measured by knowledge retention under parameter space perturbations).

4.2 Effective Rank: Measures, Generalization, and Robustness

The notion of effective rank provides a real-valued extension of matrix rank, quantifying the number of significant singular directions in a weight matrix, and has been widely adopted for characterizing NNs' intrinsic dimensionality of learned representations (Roy & Vetterli, 2007; Huh et al., 2021).

A line of work has identified what is known as the *low-rank simplicity bias* in deep NNs (Huh et al., 2021). Huh et al. demonstrated that deeper NNs are inductively biased to find solutions with lower effective rank embeddings, a bias present both at initialization and after training, resilient to hyper-parameters and learning methods, and conjectured to exist because the volume of functions that map to low effective rank embeddings increases with depth; on natural data these low-rank solutions are often the ones that generalize well (Huh et al., 2021).

The relationship between effective rank and robustness, however, is more nuanced and context-dependent. Langenberg et al. showed that adversarial training promotes simultaneously low-rank and sparse weight structure, and that explicitly promoting low rank via nuclear norm regularization improves adversarial robustness (Langenberg et al., 2019). Low effective rank is therefore associated with higher robustness to *input-space perturbations* at fixed parameters. This complements the low-rank simplicity bias of standard training, although the mechanisms differ: the low-rank solutions reached by gradient descent arise from implicit regularization, whereas those induced by adversarial training or explicit regularization are actively promoted by the objective.

Our work contributes from a distinct perspective on the relationship between effective rank and robustness. While prior studies typically characterize the properties of a trained solution with its parameters fixed, we studied the model robustness under noise-injection-induced parameter updates. We find high-entropy NNs exhibit systematically higher effective rank in attention and MLP layers and substantially stronger retention than the lower-rank AdamW-Goldilocks states. These two findings do not conflict as they address different axes: while low effective rank may benefit the robustness of a fixed function to input perturbations, high rank is associated with the robustness of a learned solution under parameter-space evolution. Our results therefore suggest that the relationship between effective rank and robustness depends on what is being perturbed, inputs or model parameters, and on whether robustness is evaluated at a fixed solution or during continual learning.

4.3 Catastrophic Forgetting and Robustness to Parameter-Space Perturbations

Catastrophic forgetting, the loss of previously acquired knowledge upon learning new information, has been a central challenge since its early identification (McCloskey & Cohen, 1989), and classical approaches such as elastic weight consolidation (EWC) (Kirkpatrick et al., 2017) and synaptic intelligence (SI) (Zenke et al., 2017) penalize parameter changes that are important for previous tasks.

Recent work has begun to explore the relationship between the geometric structure of parameter space and catastrophic forgetting. Brady Steele demonstrated that forgetting in low-rank adaptation (LoRA) is governed not simply by the rank of the adapter, but also by the geometric relationship between task gradient subspaces (Steele, 2026). This finding suggests that the representational structure of the NN plays a fundamental role in determining its susceptibility to forgetting. Our noise injection experiment follows this line of study by examining how the entropy of a NN state affects its ability to retain prior knowledge when learning new information.

5 Conclusions and Discussion

5.1 Conclusions

In this work, we have extended the high-entropy advantage beyond generalization to the domain of robustness. Using noise injection experiments on modular arithmetic tasks, we demonstrated that high-entropy NNs sampled from the Boltzmann entropy landscape maintain approximately 95% test accuracy on the original task after memorizing new noisy data, whereas AdamW-trained NNs suffer from catastrophic forgetting, dropping to $\sim$77%. We term this hidden fragility behind apparent generalization the *grokked illusion*. Through singular value decomposition of NN weights, we revealed that high-entropy NNs possess significantly higher effective rank in attention and MLP layers both before and after noise injection, which agrees with the view that richer representations distribute learned information across more degrees of freedom, thereby better preserve learned functions under parameter perturbations. These findings establish that perfect generalization does not imply equal robustness, and that the Boltzmann-entropy view of a trained state can provide a principled lens for understanding both the quality and the resilience of learned representations.

5.2 Discussion

Consistent with intuition, the robustness gap between the two NN types narrows as the injected noise becomes more structurally similar to the original task, suggesting an association between the degree of representational overlap of old and new information and the extent of forgetting.

Our findings also speak to the broader literature on effective rank discussed. While previous work has shown that low-rank structures can arise from implicit regularization (standard training) or explicit compression (adversarial training), high-entropy sampling produces a third regime of high-rank representations that are intrinsically robust. High ER corresponds to a more uniform singular value spectrum, so information is distributed across a broader set of feature dimensions; we speculate that this distributed representation provides a redundancy that may allow the NN to absorb new information while retaining enough dimensions to preserve the original task, consistent with the observation that WLMD-equilibrium NNs maintain higher ER after injection despite a larger decrease. This shares the spirit of mixture-of-experts architectures, where distributing computation across parameter subspaces reduces reliance on any single computational route(Shazeer et al., 2017; Li et al., 2025). In our experiments this redundancy is not explicitly enforced during conventional training: AdamW remains a loss-minimization procedure that does not bias toward high-entropy regions, whereas WLMD explicitly biases the search toward states with larger parameter-space volume under the same constraint. This explains why the two types of NNs, despite identical weight norm and test accuracy, occupy fundamentally different regions of the parameter space.

5.3 Implications

Our findings carry several implications beyond the specific setting of modular arithmetic. Firstly, the robustness advantage of high-entropy states offers a new perspective on forgetting under rehearsal, which is the protocol studied here: the entropy of a NN's state may be a fundamental property

associated with its robustness to forgetting, opening the possibility of entropy-guided strategies that position the model in high-entropy regions which are intrinsically more robust, rather than merely regularizing parameter updates.

Secondly, the same diagnostic is relevant to large language models, whose evaluation is increasingly moving beyond static benchmarks, partly due to concerns over benchmark saturation and contamination, toward interactive agents and long-horizon tasks (Luo et al., 2025; 2026; Motwani et al., 2026; Zeng et al., 2026). In such settings, preserving acquired capabilities becomes increasingly important: performance can degrade with increasing context length and task horizon (Du et al., 2025) and continual acquisition of new knowledge remains vulnerable to forgetting (Yang et al., 2026a). Our noise-injection experiment probes this dimension by measuring not only model performance on fixed tasks, but also whether its established capabilities remain intact as it adapts to new information.

Finally, our findings add a third regime to the relationship between effective rank and robustness: previous work has studied low-rank structures that arise from implicit regularization during standard training or from explicit compression under adversarial training, where robustness is measured against input-space perturbations at fixed parameters (Langenberg et al., 2019). Here, in contrast, the parameters themselves are altered during continual learning, and high-entropy sampling produces high effective rank that, in our experiments, accompany substantially stronger retention.

5.4 Future Directions

Several directions remain for future exploration. Extending these findings to more complex tasks and to larger or more diverse architectures, and verifying how the robustness advantage scales with model size, would strengthen the generality of our conclusions. Establishing whether high effective rank is a cause or merely a correlate of robustness will require interventions that directly manipulate the spectrum or analytical studies of the entropy landscape's geometry, and the intrinsic robustness of high-entropy states motivates training procedures that bias optimization toward high-entropy regions, such as the WanD optimizer, which employs 1D WLMD to search for high-entropy solutions but faces training instability (Zhang et al., 2026).

Limitations

All comparisons involve a single sampled WLMD-equilibrium state and a single constrained AdamW state; the ten seeds in Figure 1 quantify the variability of the continual-learning stage only. Sampling further equilibrium states is prohibitively expensive in our setting (approximately 4×10^3 GPU-hours per state). Importantly, the WLMD state is not an arbitrarily selected trained solution: it is sampled from the thermodynamically defined high-entropy equilibrium region of the parameter-space landscape. From a statistical-physics perspective, it therefore represents one microscopic realization of an equilibrium macrostate, whose realizations are expected to share statistical properties at the ensemble level. Our results therefore provide evidence for an entropy-associated robustness advantage at the state level: the sampled equilibrium state consistently retains more of its original generalization than all three distinct AdamW reference states (AdamW-Goldilocks, AdamW-standard, and AdamW-full). Direct empirical validation that this robustness advantage is shared across the equilibrium ensemble will require additional independently sampled equilibrium states.

We also note that the two states compared here are not matched in training loss: the equilibrium state is less converged on the original training data than the AdamW state. However, our comparison is instead anchored at the same level of generalization, with both states achieving 100% test accuracy before continual learning. Although the training-loss difference represents a potential confounding factor, it does not obviously favor the equilibrium state. Instead, previous work has associated more consolidated and less plastic solutions with reduced forgetting Shi et al. (2021); Bian et al. (2024), where a less-converged state generally remains more susceptible to interference from subsequent updates (lower robustness). Yet our equilibrium state exhibits substantially stronger retention. Therefore, we believe the training-loss mismatch does not provide a straightforward explanation for the robustness gap observed here.

Finally, effective rank here is reported as a correlate of retention, and we do not claim that it is the cause of the observed robustness, since we do not intervene on the weight spectra.

ETHICS STATEMENT

This work involves only synthetic data generated from modular arithmetic operations and does not involve any human subjects, personal data, or sensitive information. All experiments were conducted on standard benchmarks and synthetic datasets that do not raise privacy, fairness, or safety concerns. The noise injection experiments in this study are designed to evaluate NN's robustness and do not involve any real-world applications that could cause harm. We do not foresee any direct negative societal impacts of this work, as it is a fundamental study of NN properties rather than a deployed system. All code and models used in this paper will be made publicly available upon publication to facilitate reproducibility. No large language models were used in the writing or generation of this manuscript beyond editing and polishing. All experimental design, data collection, analysis, and interpretation of results were performed entirely by the human authors. The authors assume full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

The model architecture, the training protocols, the noise injection protocol, and the effective rank analysis are described in Section 2 of the main text, and the complete experimental configuration is given in Appendix A.1. This includes the Wang-Landau Molecular Dynamics sampling setup (weight norm constraint $\|w\| = 30$, entropy bias range, histogram flatness convergence criterion, and computing infrastructure), the AdamW hyperparameters, and the fine-tuning protocol used for noise injection. All datasets are synthetically generated from modular arithmetic operations and require no special access permissions. Each noise injection condition was repeated with ten random seeds, and the reported quantities are means together with the corresponding standard deviations. All code required to reproduce the experiments, including training scripts, the noise injection protocol, and the analysis scripts, together with model checkpoints, training logs, and raw results, is provided as an anonymized supplementary code package and will be made publicly available upon publication.

AI USE STATEMENT

In this work, generative AI tools were used only for language editing and polishing of the manuscript. We have not used generative AI tools to generate research ideas, to design the experiments, to produce or analyze results, to write code, or to create any figures, tables, or data reported in this paper, and the remaining tasks with required disclosure are not applicable to this work. All AI-assisted edits were reviewed and approved by the authors, and the experimental design, data collection, analysis, and interpretation of results were performed entirely by the human authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual review of condensed matter physics*, 11(1):501–528, 2020.

Ang Bian, Wei Li, Hangjie Yuan, Chengrong Yu, Mang Wang, Zixiang Zhao, Aojun Lu, Pengliang Ji, and Tao Feng. Make continual learning stronger via C-Flat. In *Advances in Neural Information Processing Systems*, volume 37, pp. 7608–7630, 2024.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A. Huerta, and Hao Peng. Context length alone hurts

LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23281–23298, 2025.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pp. 2232–2241. PMLR, 2019.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, May 2015. URL https://arxiv.org/abs/1412.6572.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. arXiv:2103.10427, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Peter Langenberg, Emilio Rafael Balda, Arash Behboodi, and Rudolf Mathar. On the effect of low-rank weights on adversarial robustness of neural networks. arXiv:1901.10371, 2019.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness Shroff. Theory on mixture-of-experts in continual learning. In *The Thirteenth International Conference on Learning Representations (ICLR)*, Singapore, April 2025. URL https://openreview.net/forum?id=7XgKAabsPp.

Ziming Liu, Ouail Kitouni, Niklas 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, pp. 34651–34663, 2022.

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Haotian Luo, Huaisong Zhang, Xuelin Zhang, Haoyu Wang, Zeyu Qin, Wenjie Lu, Guozheng Ma, Haiying He, Yingsha Xie, Qiyang Zhou, Zixuan Hu, Hongze Mi, Yibo Wang, Naiqiang Tan, Hong Chen, Yi Fung, Chun Yuan, and Li Shen. Ultrahorizon: Benchmarking LLM-Agent capabilities in ultra long-horizon scenarios. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/61412.

Ziyang Luo, Zhiqi Shen, Wenzhuo Yang, Zirui Zhao, Prathyusha Jwalapuram, Amrita Saha, Doyen Sahoo, Silvio Savarese, Caiming Xiong, and Junnan Li. Mcp-universe: Benchmarking large language models with real-world model context protocol servers. arXiv:2508.14704, 2025.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Sumeet Motwani, Daniel Nichols, Charles London, Peggy Li, Fabio Pizzati, Acer Blake, Hasan Hammoud, Tavish McDonald, Akshat Naik, Alesia Ivanova, Vignesh Baskaran, Ivan Laptev, Ruben Glatt, Tal Ben-Nun, Phil Torr, Ameya Pandurang Prabhu, Brian Bartoldson, Bhavya Kailkhura, and Christian Schroeder de Witt. Longcot: Benchmarking long-horizon chain-of-thought reasoning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, Seoul, South Korea, July 2026. URL https://icml.cc/virtual/2026/poster/66396.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. arXiv:2201.02177, 2022.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6747–6761, 2021.

Brady Steele. Subspace geometry governs catastrophic forgetting in low-rank adaptation. arXiv:2603.02224, 2026.

Fugao Wang and David P Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical Review Letters*, 86:2050–2053, Mar 2001. doi: 10.1103/PhysRevLett.86.2050. URL https://link.aps.org/doi/10.1103/PhysRevLett.86.2050.

Cheng Yang, Xuemeng Yang, Licheng Wen, Daocheng Fu, Jianbiao Mei, Rong Wu, Pinlong Cai, Yufan Shen, Nianchen Deng, Jia Xu, Botian Shi, Yu Qiao, and Haifeng Li. Learning on the job: An experience-driven self-evolving agent for long-horizon tasks. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*, 2026a. URL https://iclr.cc/virtual/2026/10012478.

Entao Yang, Xiaotian Zhang, Yue Shang, and Ge Zhang. High-entropy advantage in neural networks' generalizability. *npj Artificial Intelligence*, 2(1):44, 2026b.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhu Zhenwei, Tianle Cai, Xinjie Chen, Zehui Chen, Jiecao Chen, Yantao Du, Xiang Gao, Jiacheng Guo, Liang Hu, Jianpeng Jiao, Xiangsheng Li, Jingkai Liu, Nishuang, Zhoufutu Wen, Ge Zhang, Kaiyuan Zhang, Xin Zhou, Jose Blanchet, Xipeng Qiu, Mengdi Wang, and Wenhao Huang. Futurex: An advanced live benchmark for LLM agents in future prediction. In *The Fourteenth International Conference on Learning Representations (ICLR)*, Rio de Janeiro, Brazil, April 2026. URL https://iclr.cc/virtual/2026/poster/10006496.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. Pmlr, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations (ICLR)*, Toulon, France, April 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Xiaotian Zhang, Yue Shang, Entao Yang, and Ge Zhang. Is grokking a computational glass relaxation? *Advances in Neural Information Processing Systems*, 38:101465–101489, 2026.

A Supplementary Material for The Grokked Illusion: True Equilibrium Mitigates Catastrophic Forgetting

A.1 Experimental Setting

A.1.1 WLMD Sampling Configuration

The high-entropy NN used in this study was obtained via Wang-Landau Molecular Dynamics (WLMD) sampling, following the protocol established in (Zhang et al., 2026). The sampling was performed under the same weight norm constraint ($\|w\| = 30$) and entropy bias range ($\ln(\mathcal{L}_{\text{train}}) \in [-20, 1]$) as described in the main text. The simulation was run with 8 parallel WLMD processes on 4 GeForce RTX 4090 GPUs, requiring approximately 100 GPU hours per 1 million WLMD update epochs (Zhang et al., 2026).

Convergence of the entropy landscape was assessed by monitoring the histogram flatness criterion. We observed that after approximately 40 million WLMD update epochs, the entropy landscape exhibited no substantial further change, indicating that the sampling had reached a converged regime. From this converged regime, we selected one representative high-entropy NN state for the noise injection experiments reported in the main text.

The network architecture used in the WLMD sampling is identical to that described in the main text: a single-layer Transformer with $d_{\text{vocab}} = 68$, $d_{\text{model}} = 128$, $d_{\text{mlp}} = 512$, $d_{\text{head}} = 32$, and 4 attention heads, totaling approximately 2.15×10^5 parameters.

A.1.2 AdamW Training and Noise Injection Configuration

The AdamW-goldilocks NNs used in this study were trained under the same weight norm constraint ($\|w\| = 30$) enforced via linear scaling after each parameter update, following the protocol established in (Liu et al., 2023). The training employed an AdamW optimizer with learning rate 1×10^{-2}, weight decay 1.0, and full-batch gradient updates for 120 epochs.

For the noise injection experiments, the vocabulary was expanded from 68 to 136 tokens, with the original task tokens occupying indices $0, \ldots, 67$ and the new noise tokens occupying indices $68, \ldots, 104$. Fine-tuning on the mixed dataset (original training data + 500 noise samples) was performed using AdamW with learning rate 1×10^{-4}, full-batch updates, and stopping at $> 99.8\%$ training accuracy on the mixed dataset.

Two additional AdamW variants were examined as control experiments:

- **AdamW-standard**: Trained with identical hyperparameters as AdamW-goldilocks, except that the weight norm constraint ($\|w\| = 30$) is *not* enforced.
- **AdamW-full**: Trained with identical hyperparameters as AdamW-goldilocks, including the weight norm constraint, except that the training set comprises *both* the original training and test samples of the modular arithmetic task.

All subsequent noise injection procedures for these two variants follow the same protocol described above. All code necessary to reproduce the experiments, including training scripts, noise injection protocols, and analysis scripts, is provided in the supplementary code package. The package also includes all model checkpoints, training logs and results.

A.2 Code and Data Availability

All code necessary to reproduce the experiments will be made publicly available upon publication. The datasets are synthetically generated and do not require special access permissions.

A.3 Supplementary Experimental Results

This section presents supplementary experimental results that are referenced but not shown in the main text due to space constraints.

A.3.1 Complete SVS Plots for WLMD-Equilibrium and AdamW-Goldilocks

Figures 2–7 present the complete Singular Value Spectrum (SVS) plots for all nine layers of the AdamW-goldilocks and WLMD-equilibrium NNs under each of the three noise conditions. In each subplot, the blue curve denotes the pre-injection spectrum and the red curve denotes the post-injection spectrum. The subplot title reports the effective rank (ER) before and after injection, and the cosine similarity between the pre- and post-injection weights.

Figures 2–4 correspond to AdamW-goldilocks under random noise, $x^2+y \bmod 37$, and $x+y \bmod 37$, respectively. Figures 5–7 correspond to WLMD-equilibrium under the same three noise conditions, in the same order.

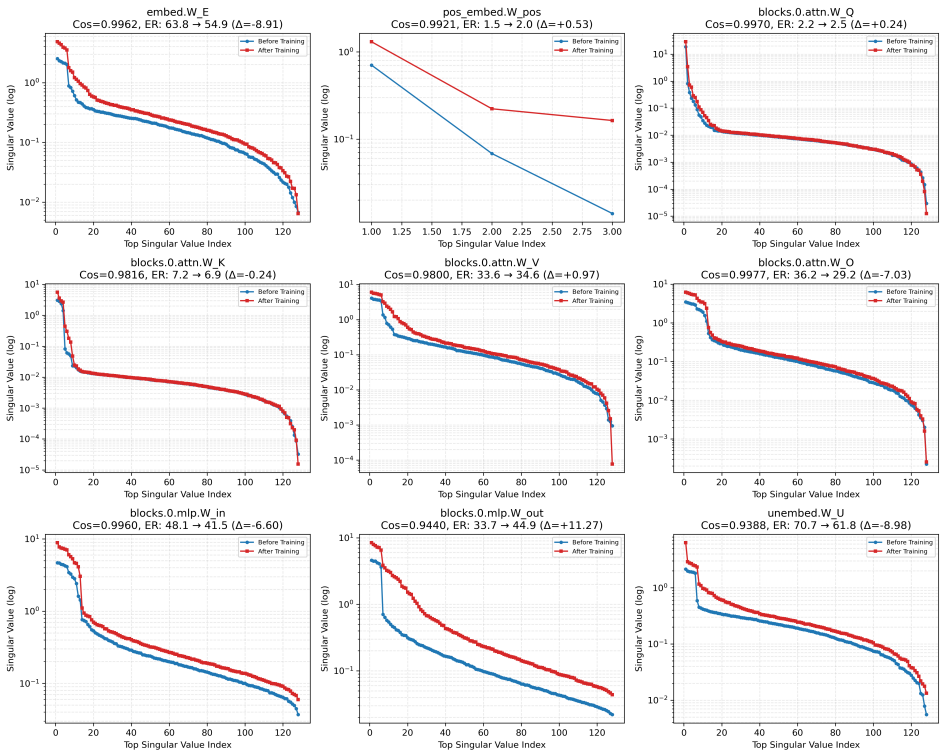


Figure 2: Complete SVS plots for AdamW-goldilocks under random noise.

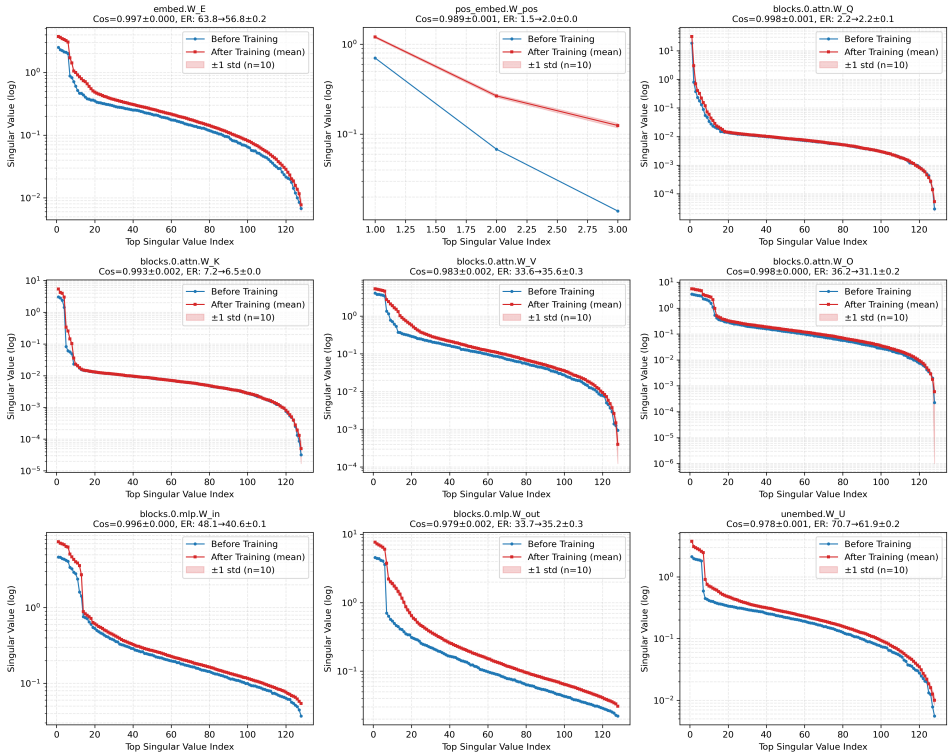


Figure 3: Complete SVS plots for AdamW-goldilocks under $x^2 + y \bmod 37$ noise.

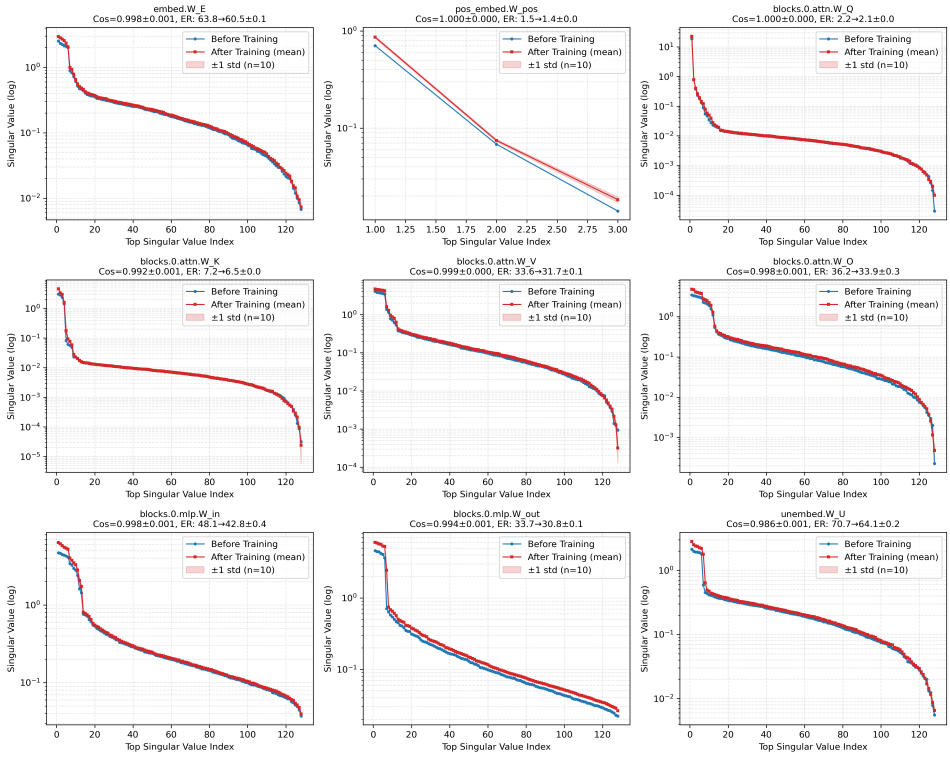


Figure 4: Complete SVS plots for AdamW-goldilocks under $x + y \bmod 37$ noise.

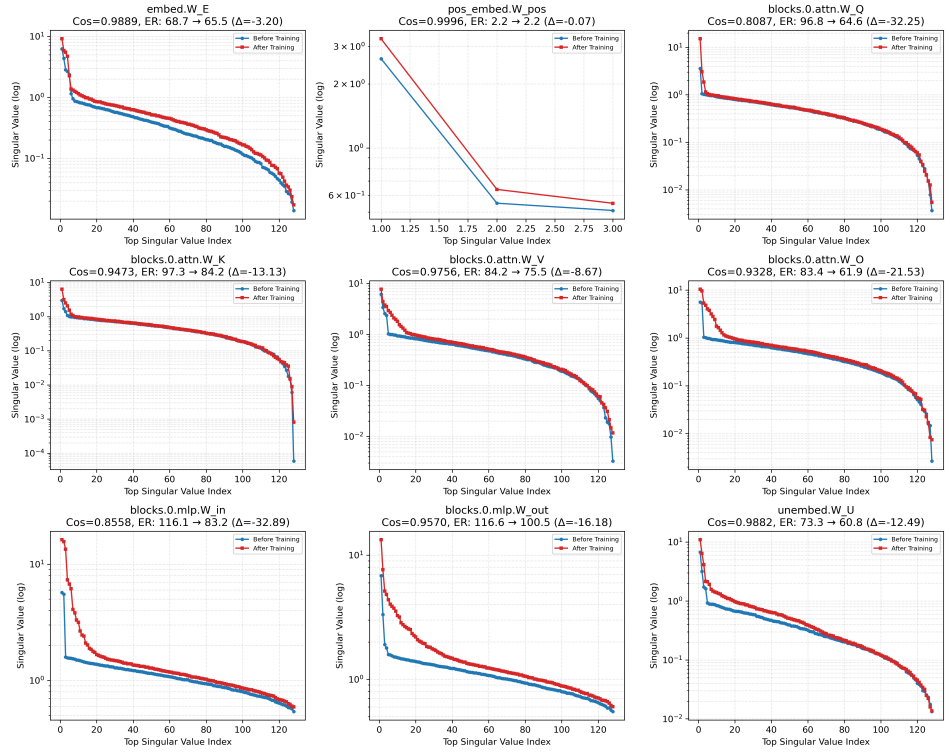


Figure 5: Complete SVS plots for WLMD-equilibrium under random noise.

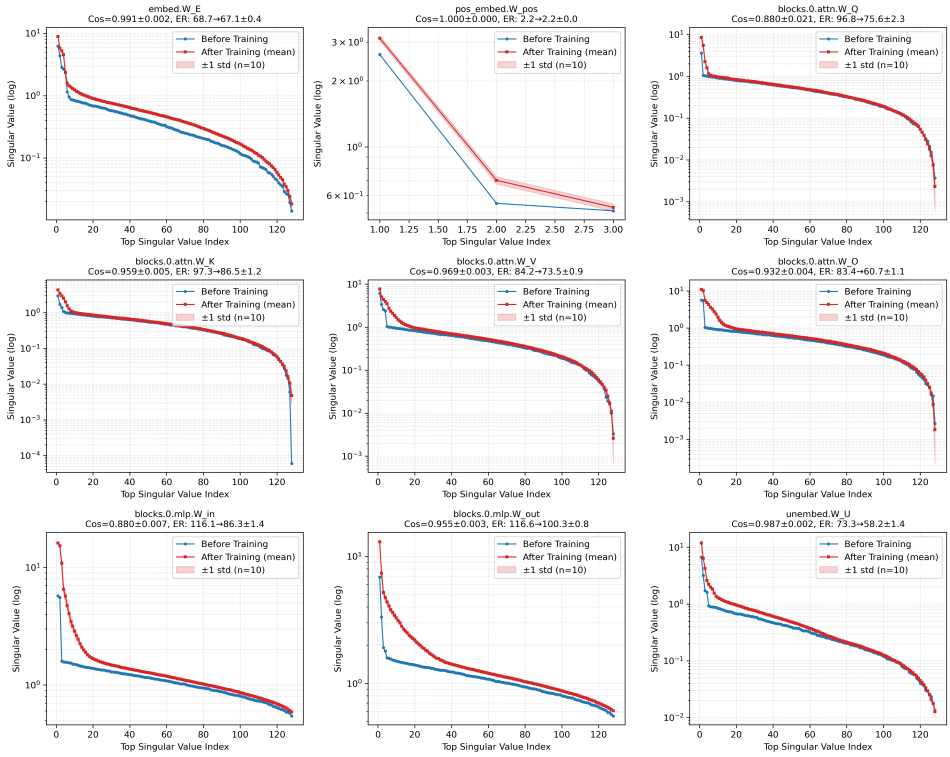


Figure 6: Complete SVS plots for WLMD-equilibrium under $x^2 + y \bmod 37$ noise.

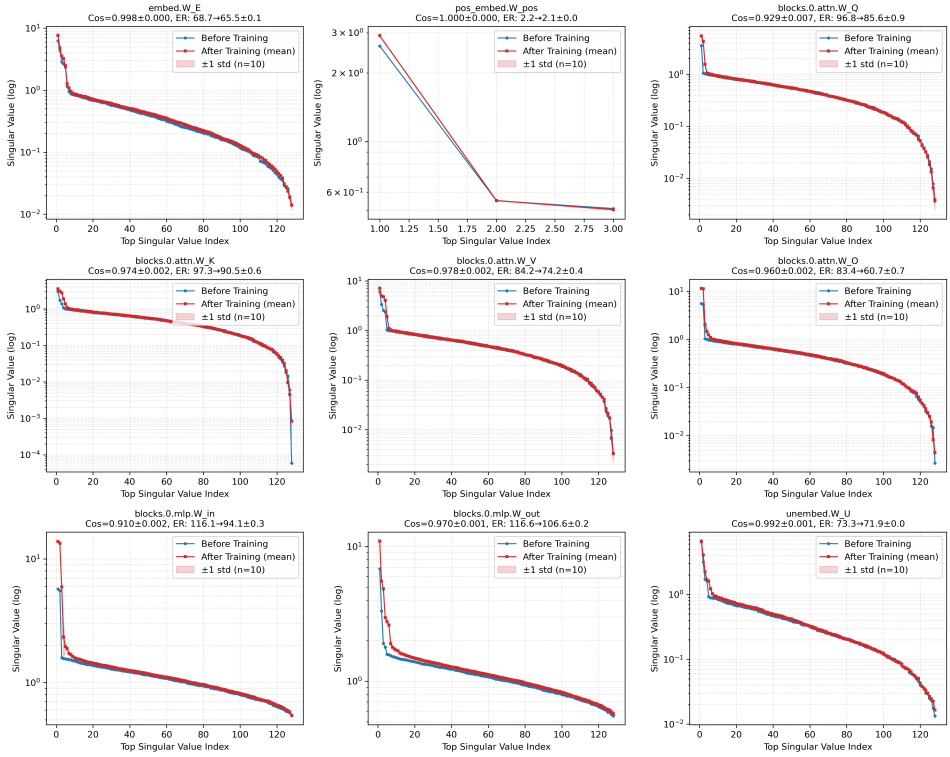


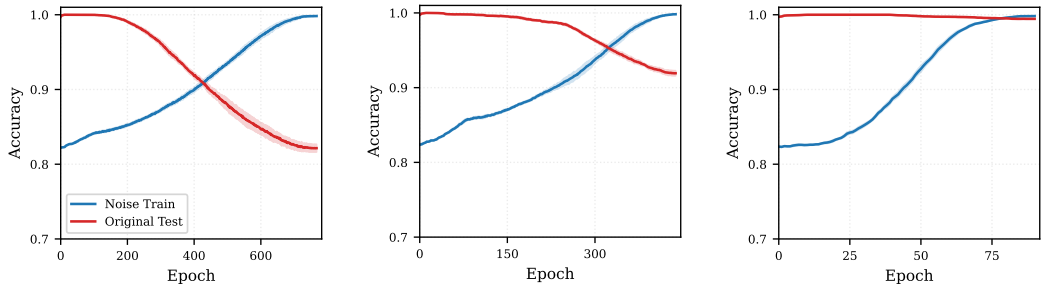
Figure 7: Complete SVS plots for WLMD-equilibrium under $x + y \bmod 37$ noise.

A.3.2 RESULTS FOR ADAMW-STANDARD

AdamW-standard refers to NNs trained with standard AdamW optimization without the weight norm constraint ($\|w\| = 30$) enforced during training. All other hyperparameters are identical to those of AdamW-goldilocks.

Figure 8 shows the noise injection training dynamics for AdamW-standard under the three noise types. Although AdamW-standard achieves $> 99.8\%$ training accuracy on the mixed dataset in all cases, its original-task test accuracy drops varies: dropping to approximately 82% under random noise, 92% under $x^2 + y \bmod 37$, and 98% under $x + y \bmod 37$. The robustness of AdamW-standard is better than that of AdamW-goldilocks, which may be attributed to the absence of the strict weight norm constraint that forces the NN into a more constrained region of the parameter space. Nevertheless, AdamW-standard still exhibits a significant robustness gap compared to WLMD-equilibrium, confirming that the stability advantage of high-entropy networks is not merely an artifact of the weight norm constraint but rather a fundamental property of high-entropy states.

Figures 9–11 present the complete SVS plots for AdamW-standard under the three noise conditions.



(a) Random noise (b) $x^2 + y \bmod 37$ (c) $x + y \bmod 37$

Figure 8: Training dynamics of AdamW-standard under three noise types. Red lines indicate test accuracy on the original task; blue lines indicate training accuracy on the mixed dataset. Shaded areas represent ± 1 standard deviation over 10 random seeds.

A.3.3 RESULTS FOR ADAMW-FULL

AdamW-full refers to NNs trained with standard AdamW optimization on the *full* dataset, which includes both the original training and test samples of the modular arithmetic task, under the same weight norm constraint ($\|w\| = 30$) as AdamW-goldilocks. This control experiment is designed to rule out the concern that the WLMD-equilibrium NN's robustness advantage might stem from test-data exposure during the entropy landscape sampling process.

Figure 12 shows the noise injection training dynamics for AdamW-full under the three noise types. The results are nearly identical to those of AdamW-standard, with original-task test accuracy dropping to approximately 82% under random noise, 93% under $x^2 + y \bmod 37$, and 99% under $x + y \bmod 37$, which is still inferior to WLMD-equilibrium NN. This result indicates that the robustness advantage of the WLMD-equilibrium NN is not explained by exposure to the test data, and is consistent with the high-entropy nature of the sampled states.

Figures 13–15 present the complete SVS plots for AdamW-full under the three noise conditions.

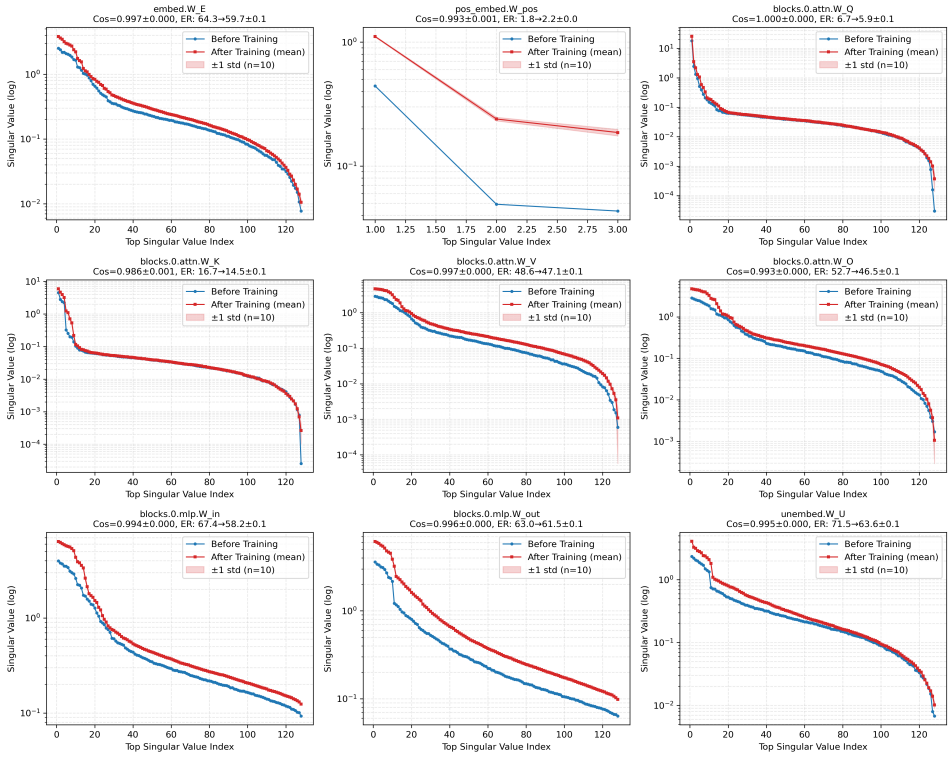


Figure 9: Complete SVS plots for AdamW-standard under random noise.

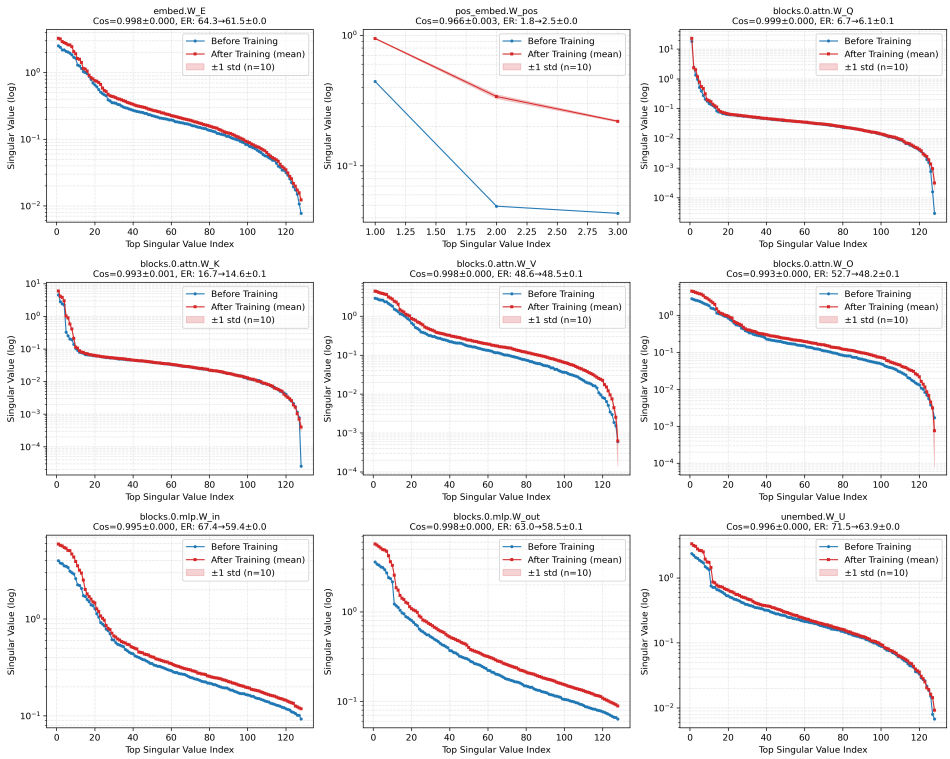


Figure 10: Complete SVS plots for AdamW-standard under $x^2 + y \bmod 37$ noise.

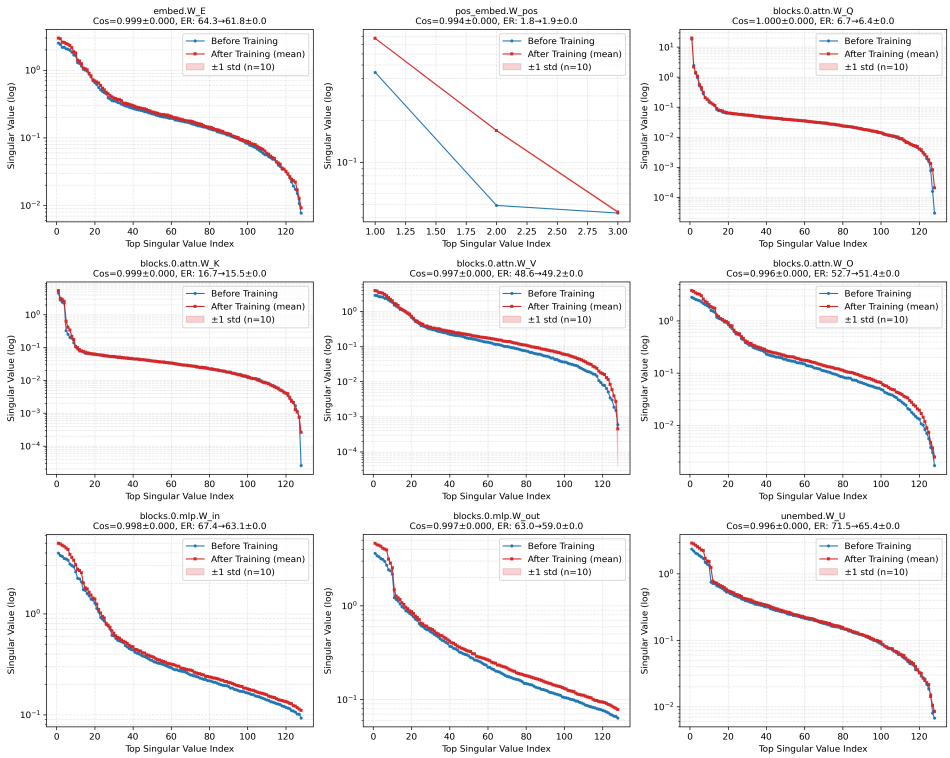
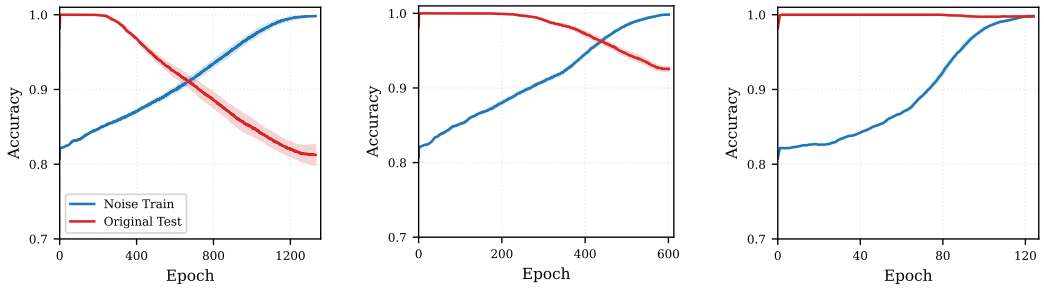


Figure 11: Complete SVS plots for AdamW-standard under $x + y \bmod 37$ noise.



(a) Random noise (b) $x^2 + y \bmod 37$ (c) $x + y \bmod 37$

Figure 12: Training dynamics of AdamW-full under three noise types. Red lines indicate test accuracy on the original task; blue lines indicate training accuracy on the mixed dataset. Shaded areas represent ± 1 standard deviation over 10 random seeds.

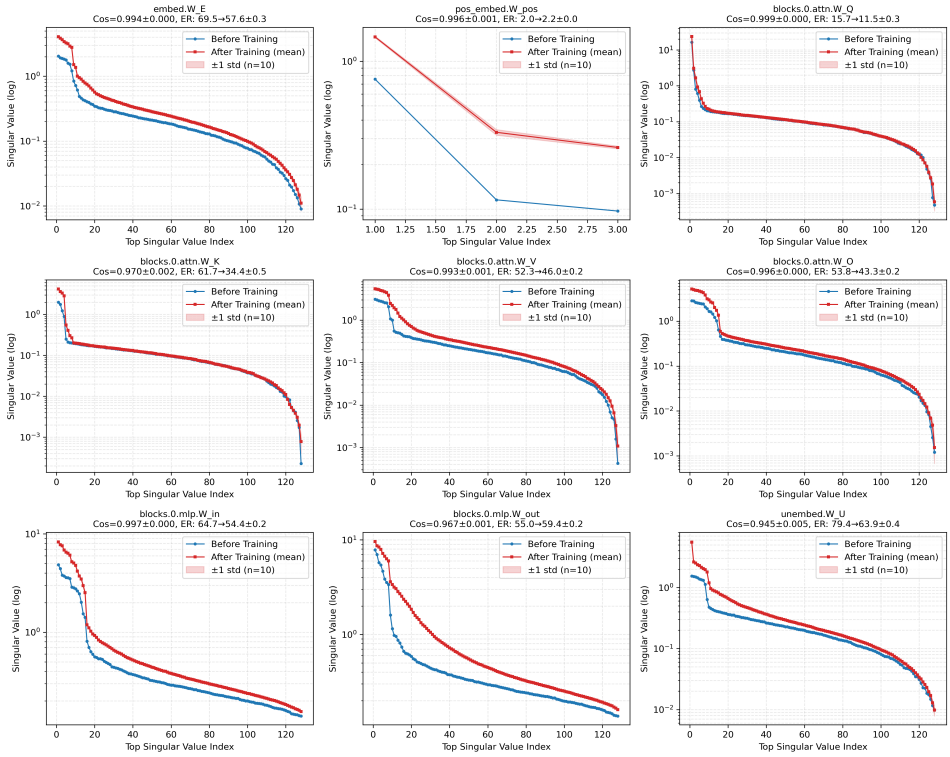


Figure 13: Complete SVS plots for AdamW-full under random noise.

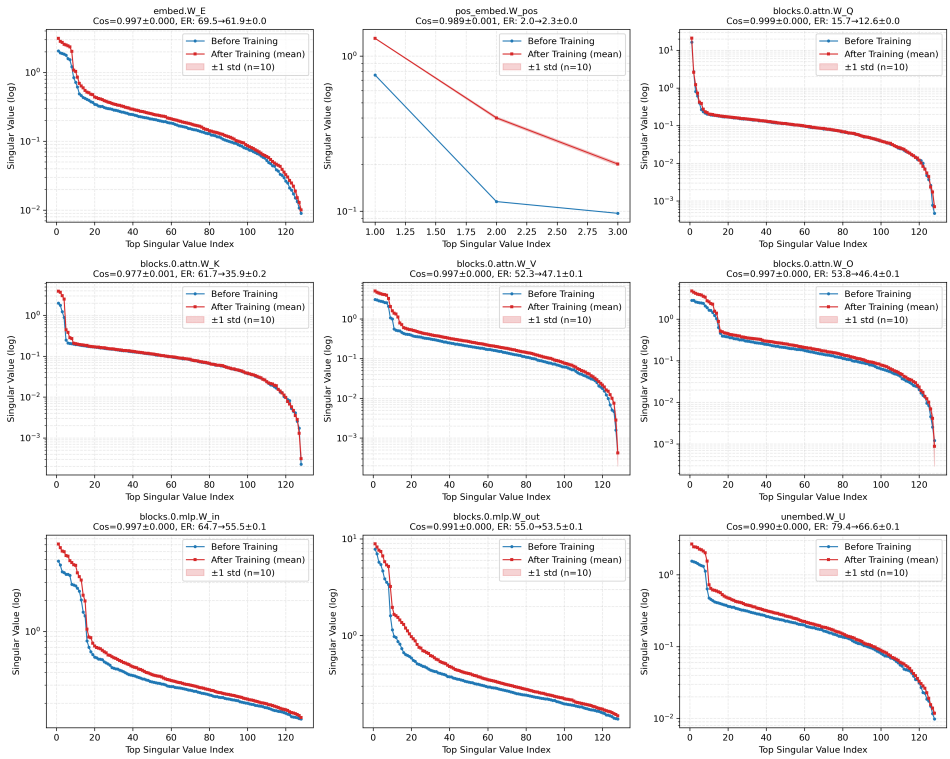


Figure 14: Complete SVS plots for AdamW-full under $x^2 + y \bmod 37$ noise.

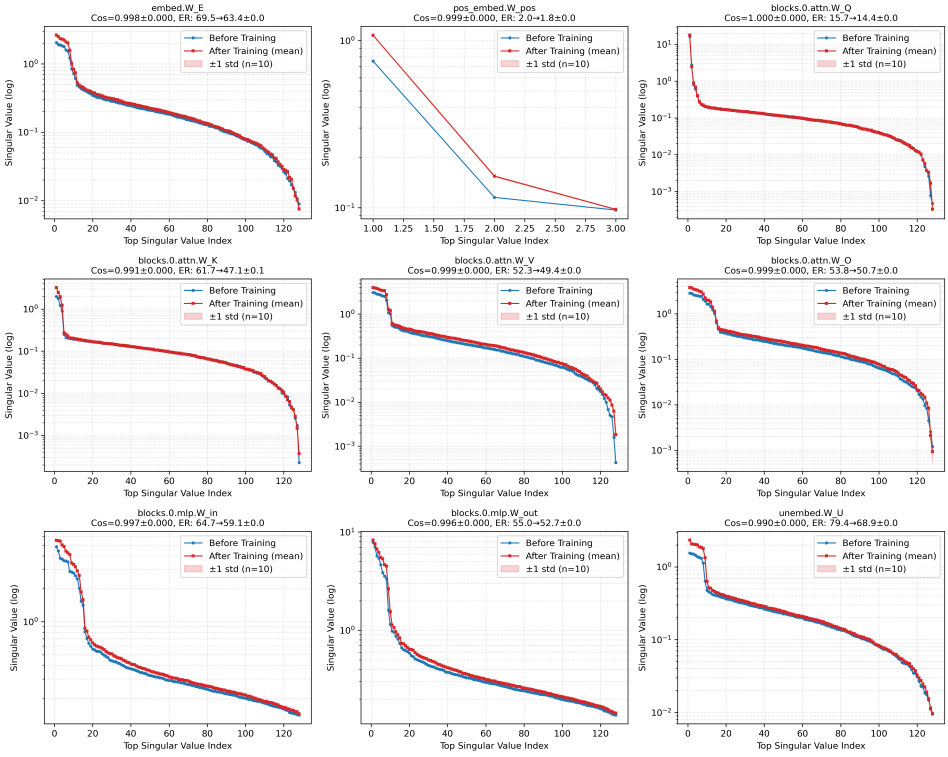


Figure 15: Complete SVS plots for AdamW-full under $x + y \bmod 37$ noise.