Journal of High School Science

Investigating the impact of low to moderate level label noise on Neural Network grokking dynamics

Singavarapu H

Submitted: September 21, 2025, Revised: version 1, November 18, 2025, version 2, November 24, 2025.
Accepted: November 29, 2025

Abstract

Neural networks trained on algorithmic tasks, given specific model architecture features, can sometimes exhibit grokking—a phenomenon where perfect generalization emerges after a lag time subsequent to the network having memorized the training data. While previous research has focused on weight decay and dataset size as factors influencing grokking, the role of label noise remains unexplored. This study systematically investigated whether controlled label corruption modulated grokking dynamics in neural networks given the task of learning modular addition. We trained 155 neural networks on the task (a + b) mod 97 with varying low to moderate levels of random label noise from 0% to 15% in 0.5% increments applied to the training dataset only. Statistical validation using exponential regression with confidence intervals ($R^2 > 0.85$ for all metrics) and Levene's test ($p < 0.001$) revealed that the generalization gap and time to grok scaled exponentially with noise level. Contrary to expectations that small amounts of noise might act as beneficial regularization, we found that even minimal label corruption (0.5%) measurably degraded grokking dynamics, with effects compounding exponentially as noise increased. Robustness validation across three independent random train-test splits and five hyperparameter configurations confirmed that exponential scaling persisted despite variations in data partitioning and optimization settings. These findings, while specific to our experimental configuration, provide insight into how label noise fundamentally hinders the ability of neural networks to discover generalizing solutions after periods of overfitting in specific algorithmic tasks.

Keywords

Grokking, Delayed generalization, Label noise, Neural Networks, Modular addition, Overfitting, Training dynamics, Exponential scaling, Test accuracy, Training accuracy, Test loss, Training loss

Havish Singavarapu, Newport High School,4333 Factoria Boulevard SE, Bellevue, Washington, 98006, USA.
havish279@gmail.com

https://doi.org/10.64336/001c.154269

1. Introduction

Grokking is described as a phenomenon where overfit neural networks exhibit delayed generalization after an extended period of perfect training accuracy. The conundrum of grokking represents one of the most intriguing discoveries in modern deep learning research. First documented by Power et al. (1) in transformer models learning modular arithmetic, grokking challenges our fundamental understanding of how neural networks learn and generalize. While initial investigations focused on clean, deterministic datasets, real-world applications invariably contend with noisy labels, raising critical questions about the robustness of grokking dynamics under imperfect training conditions. This paper investigates how label noise influences the grokking phenomenon in Multi-Layer Perceptrons (MLPs), providing insights into both the mechanisms underlying delayed generalization and the practical implications for training neural networks on noisy data.

The intersection of grokking and label noise represents unexplored territory in deep learning theory. Previous work established that grokking occurs reliably in algorithmic tasks with high weight decay (2) and that the phenomenon relates to the simplicity bias of neural networks (3). Separately, the machine learning community has extensively studied learning under label noise (4, 5), demonstrating that neural networks can be surprisingly robust to corrupted labels through implicit regularization mechanisms (6, 7). However, how these two phenomena interact—whether label noise accelerates, delays, or fundamentally alters grokking dynamics—remains an under-researched area which has significant theoretical and practical implications.

In broader deep learning contexts, controlled noise has proven beneficial for generalization. Dropout (8), which randomly deactivates neurons during training, has become a standard regularization technique, reducing overfitting by up to 2% on ImageNet tasks. Label smoothing (9), which deliberately adds noise to one-hot encoded targets, improves model calibration and generalization in neural machine translation and image classification. Data augmentation techniques that add noise to inputs have shown consistent improvement across vision tasks (10). The noisy student training paradigm (11) achieves state-of-the-art results by training on pseudo-labels that inherently contain errors. Given these successes, investigating whether similar noise-based benefits extend to the grokking phenomenon represents a natural and important research direction.

1.1 Related work

Our work on grokking under label noise connects several research areas: the grokking phenomenon itself, learning with noisy labels, and the intersection of regularization and generalization in deep learning. We review each area and position our contributions.

1.1.1 *Grokking discovery and initial characterization*

The grokking phenomenon was first documented in seminal work on transformer models, which observed that neural networks trained on small algorithmic datasets could exhibit perfect generalization long after

completely overfitting the training data. Their influential work established modular arithmetic tasks as the canonical testbed for studying grokking, demonstrating that models required millions of training steps to transition from memorization to generalization. This work introduced the terminology and experimental framework that subsequent research, including ours, builds upon.

Liu et al. (12) provided the first theoretical framework for understanding grokking through their "effective theory of representation learning." They identified four distinct learning phases - comprehension, grokking, memorization, and confusion—and showed that grokking occurs in a "Goldilocks zone" between memorization and confusion. Their phase diagram approach inspired our systematic exploration of noise levels. Unlike their work focusing on weight decay and learning rate interactions, we introduce label noise as a new factor of variation.

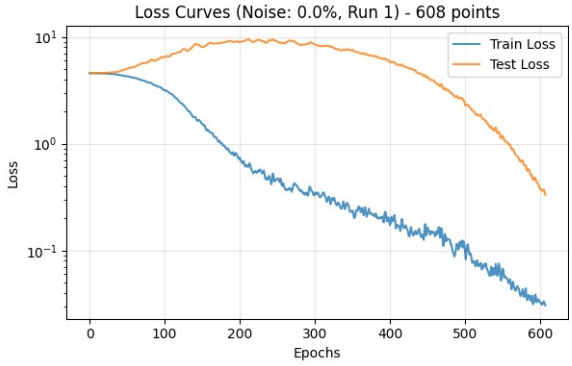

Figure 1. Graph of the training loss (blue) and test loss (orange) curves at 0% noise

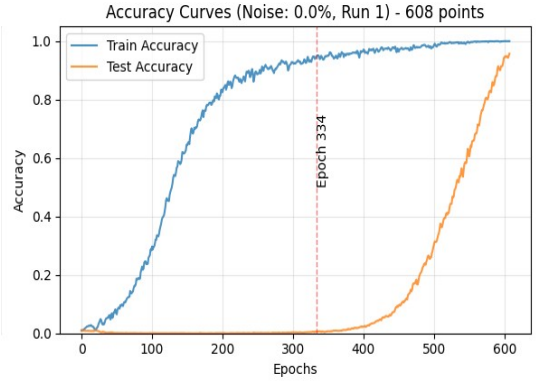

Figure 2. Graph of the training accuracy (blue) and test accuracy (orange) curves at 0% noise

1.1.2 *Mechanistic understanding*

Detailed mechanistic interpretability analysis has reverse-engineered how transformers solve modular addition through Fourier transforms and trigonometric identities. Progress measures have been introduced, including "restricted loss" and "excluded loss", that track the emergence of generalizing circuits. While these measures focused on understanding the learned algorithm, our work examines how noise affects the learning dynamics leading to these algorithms. Varma et al. (13) proposed that grokking arises from competition between memorizing and generalizing circuits, with weight decay favoring more efficient generalizing solutions. This circuit efficiency hypothesis complements our findings that noise may prevent the formation of memorizing circuits, thereby accelerating the discovery of generalizing solutions.

1.1.3 *Theoretical perspectives*

Kumar et al. (14) argued that grokking emerges from the transition between "lazy" and "rich" training regimes, showing that grokking can occur without explicit regularization in certain settings. They demonstrated this on polynomial regression with two-layer networks. Our work extends this conjecture by showing that label noise can induce similar regime transitions even in deeper networks where weight decay is typically required.

Lyu et al. (15) established rigorous conditions for grokking, demonstrating that large initialization and small weight decay were essential in their studied settings. They introduced the concept of "misgrokking"—where validation accuracy initially improves and then degrades. We build on their analysis by showing that label noise can modulate these conditions, potentially preventing misgrokking while promoting standard grokking.

1.1.4 *Classical approaches to learning with noise*

The problem of learning with label noise has been extensively studied in machine learning. Natarajan et al. (16) provided theoretical bounds for learning with random classification noise, showing that uniform label noise acted as a form of regularization. Patrini et al. (17) developed loss correction methods for deep neural networks under label noise. While these works focus on achieving robustness despite noise, we investigate whether noise may enhance generalization through grokking.

Prior work famously showed that deep networks can perfectly memorize random labels, raising questions about why they generalize well on clean data. This work highlighted the importance of implicit regularization in deep learning. Our research provides a complementary perspective: even when networks memorize noisy labels, they can subsequently discover generalizing solutions through extended training.

1.1.5 *Extension to real-world datasets*

Grokking has been successfully induced on MNIST and IMDb datasets by manipulating initialization scale and weight decay, demonstrating that grokking is not limited to algorithmic tasks. The "LU mechanism" (mismatch between L-shaped training loss and U-shaped test loss curves) underlies grokking across domains. Our noise-based approach

provides an alternative method for inducing grokking that may be more practically applicable. Miller et al. (18) showed that grokking occurs in Gaussian Process regression and other non-neural models, suggesting that it is a general phenomenon in machine learning rather than specific to neural networks. This broader perspective supports our investigation of noise as a general mechanism for inducing grokking.

1.1.6 *Grokking in language models*

Recent work has observed grokking-like phenomena in large language models. Millidge et al. (19) documented delayed generalization in transformer language models on various tasks. Thilak et al. (20) found that language models exhibited staged learning, with sudden improvements in performance after extended training plateaus. These observations suggest that our noise-based grokking mechanisms may be relevant for understanding training dynamics in practical applications.

1.1.7 *Quantifying grokking*

Various metrics have been proposed to identify and quantify grokking, including the gap between training and validation accuracy, the "Relative Quantity of Interest" (RQI) metric, and Fourier-based progress measures. Our contribution of the "Grokking Index"—the integrated area between accuracy curves—provides a scalar summary that captures both the duration and severity of the grokking phenomenon, enabling systematic comparison across experimental conditions.

Notsawo et al. (21) proposed spectral signatures for predicting grokking without full training, using Fourier analysis of learning curves. Our complementary approach uses integration-based measures that require full training but provide a more comprehensive characterization of the grokking trajectory.

1.1.8 *Noise as a regularization tool*

Beyond robustness to naturally occurring label noise, the machine learning community has discovered that deliberately adding noise can improve generalization. As discussed in section 1. (Introduction), dropout reduces overfitting, label smoothing improves accuracy and calibration, data augmentation boosts performance, and noisy student training achieves state-of-the-art results. These successes suggest that noise serves as implicit regularization, preventing overfitting to training data specifics. However, whether these benefits extend to grokking—a phenomenon involving delayed rather than immediate generalization—remains unexplored.

1.2 Gaps and our contributions

Despite extensive research on grokking and learning with noisy labels, several gaps remain. No previous study has systematically characterized how label noise affects grokking dynamics, mapped how varying noise levels affect grokking behavior, or developed unified metrics that capture the full trajectory of delayed generalization. Our work addresses these gaps by demonstrating that label noise reliably delays grokking, providing fine-grained analysis across 31 noise levels with statistical validation, and introducing the Grokking Index and Loss Index as comprehensive scalar measures. This positions our work at the intersection of grokking research and noisy label learning,

opening new directions for understanding and controlling delayed generalization in neural networks.

1.3 Concurrent and future directions

Recent concurrent work continues to expand our understanding of grokking. Humayun et al. (22) demonstrated that grokking occurred in vision transformers on image classification. Chen et al. (23) documented "sudden drops" in loss during online learning, potentially related to grokking. These developments suggest that our noise-based approach may have broad applicability beyond algorithmic tasks. Future research directions inspired by our work might include extending this analysis to other algorithmic operations, investigating training strategies that mitigate the exponential cost of learning from noisy labels in grokking settings, and theoretical analysis of how noise modifies the effective loss landscape during grokking.

2. Materials and Methods

2.1 Algorithmic task setup and dataset construction

We investigated grokking on the modular addition task (a + b) mod p, where p = 97 was chosen as a prime number to ensure that the operation formed a group. This task required the model to learn the mapping (a, b) → c, where c is (a + b) mod 97. Modular addition is the standard task for studying grokking, first used by Power et al. (2022) in their seminal paper (1).

Table 1. Dataset generation

Component	**Specification**
Input space	{(i, j) | i, j ∈ {0, 1, ..., 96}}
Target outputs	{(i + j) mod 97 | for all input pairs}
Total examples	p^2 = 9,409 unique input pairs
Data shuffling	All pairs randomly shuffled before splitting

Table 2. Train-test split configuration

Parameter	**Value**
Training fraction	40% (3,764 examples)
Test fraction	60% (5,645 examples)
Split method	Random uniform sampling without replacement
Consistency	Same split maintained throughout each experimental run

The unique 40/60 train-test split was used to capture the grokking phenomenon more clearly and in a more computationally feasible timeframe. The delayed generalization was more evident with this split, as generalization was harder with only 40% of the data allocated for training.

Table 3. Label noise injection protocol

Parameter	Specification
Noise levels	31 levels from 0% to 15% in 0.5% increments
Selection method	Select $\lfloor \alpha \times \lvert \text{train set} \rvert \rfloor$ training examples uniformly at random
Label replacement	Replace correct labels with random values from {0, 1, ..., 96}
Test set	Maintained clean (no noise) for accurate generalization assessment
Noise persistence	Same corrupted labels used across all epochs for consistency

2.1.1 *Model architecture*

We employed a deep Multi-Layer Perceptron (MLP) with embedding layers. Details of the different components of the architecture are shown in Table 4.

Table 4. Network architecture

Component	Specification
Embedding dimension	128
Vocabulary size	97 (one per modular value)
Embedding structure	Two separate embeddings for first and second operands
Input processing	Concatenate embeddings: [Embedding(x_1) || Embedding(x_2)]
Dimension after concat	256
Hidden Layer 1	Linear(256 → 512) + ReLU
Hidden Layer 2	Linear(512 → 512) + ReLU
Hidden Layer 3	Linear(512 → 512) + ReLU
Output Layer	Linear(512 → 97)

2.1.2 *Weight initialization*

Linear layers used Xavier uniform initialization with gain = 0.5, biases were initialized to zeros, and embeddings followed a normal distribution N(0, 0.02). Smaller initialization promoted grokking behavior.

2.1.3 *Regularization*

Dropout was disabled (set to 0.0) for cleaner grokking dynamics, while weight decay was applied through the optimizer as described below.

2.1.4 *Training configuration*

The values for the different hyperparameters used during the training configuration are shown in Table 5.

Table 5. AdamW optimizer hyperparameters

Parameter	**Value**
Optimizer	AdamW
Base learning rate	5×10^{-3}
Betas	(0.9, 0.999)
Epsilon	1×10^{-8}
Weight decay	1.0 (strong regularization to induce grokking)

2.1.5 *Learning rate schedule*

The learning rate followed a warmup-decay schedule. During warmup (epochs 0-100), the learning rate increased linearly from 0 to the base rate: $\lambda(\text{epoch}) = (\text{epoch} + 1) / 100$. After warmup (epochs > 100), exponential decay was applied: $\lambda(\text{epoch}) = 0.99995^{(\text{epoch} - 100)}$.

Table 6. Training parameters

Parameter	**Value**
Maximum epochs	50,000
Batch size	256
Loss function	Cross-entropy loss
Gradient clipping	L2 norm clipped at 1.0 for stability
Device	CUDA GPU

2.2 Experimental design

Table 7. Main experiment configuration

Parameter	**Value**
Noise levels	31 distinct values {0%, 0.5%, 1.0%, ..., 15.0%}
Repetitions	5 independent runs per noise level
Total experiments	155 training runs
Random seeds	Different initialization for each run

2.2.1 *Split validation experiment*

To verify that exponential scaling was robust to train-test partition choice, we created three independent random splits of the 9,409 modular addition pairs. Each split maintained the 40/60 train-test ratio but differed in which specific examples were assigned to training versus test sets. For each split, we trained 5 networks with different random initializations at noise levels of 0%, 5%, 10%, and 15% (60 total runs). The first 10 examples from each split are documented in section 7 (Supplementary materials) for verification of independence.

Table 8. Hyperparameter ablation configurations

Configuration	Weight Decay	Learning Rate	Notes
Original	1.0	5×10^{-3}	Baseline
WD_0.5	0.5	5×10^{-3}	Lower reg.
WD_2.0	2.0	5×10^{-3}	Higher reg.
LR_1e-3	1.0	1×10^{-3}	Lower LR
LR_1e-2	1.0	1×10^{-2}	Higher LR

Each configuration was tested at noise levels of 0%, 5%, 10%, and 15% with 5 independent runs per condition (100 total runs).

2.2.2 *Metrics and progress measures*

Tables 9 and 10 provide descriptions of indices and specifications.

Table 9. Grokking index definition

Attribute	Specification
Definition	Area between training and test accuracy curves
Significance	Quantifies delayed generalization; measures memorization without understanding
Integration start (α)	First epoch where training accuracy ≥ 95%
Integration end (β)	First epoch where test accuracy ≥ (training accuracy − 5%)
Calculation	GI = $\int_{\alpha}^{\beta}$ (TA − tA) de (trapezoidal integration)

Table 10. Loss index definition

Attribute	Specification
Definition	Area between test and training loss curves
Significance	Cumulative "confidence gap"; model uncertainty on test vs training data
Integration bounds	Same as Grokking Index (α to β)
Calculation	LI = $\int_{\alpha}^{\beta}$ (tL − TL) de (trapezoidal integration)

The Grokking Index captures the duration of time spent in the overfitting regime, the severity (magnitude of the generalization gap), and the speed of eventual recovery.

2.2.3 *Analysis methods*

2.2.3.1 Grokking detection

A run exhibited grokking if the training accuracy reached ≥ 95% and test accuracy subsequently reached within 5% of training accuracy.

2.2.3.2 Statistical analysis

We calculated 95% confidence intervals for summary statistics at each noise level using the t-distribution (n=5 per level). To assess homogeneity of variance across noise levels, we applied Levene's test. Exponential regression models were fit using non-linear least squares, with parameter uncertainty estimated from the

covariance matrix. All statistical analyses were performed using SciPy and NumPy in Python.

2.2.3.3 Visualization

Training dynamics were visualized using loss curves (log scale) and accuracy curves with 95% threshold markers at epoch-by-epoch resolution. Summary visualizations included noise level versus final accuracies (with error bars), noise level versus Grokking/Loss indices, and correlation scatter plots between indices.

2.3 Theoretical framework

2.3.1 *Noise-induced regularization hypothesis*

We investigated whether label noise acted as implicit regularization that prevented early memorization of training data and forced the discovery of robust generalizing features.

2.3.2 *Implementation details*

Experiments were implemented in PyTorch 2.0+ with Matplotlib 3.5+ for visualization, NumPy 1.21+ and Pandas 1.3+ for data handling, and tqdm for progress tracking.

2.3.3 *Experimental variations*

While systematic ablation of architectural and optimization hyperparameters could provide additional mechanistic insights into the noise-grokking relationship, we leave comprehensive exploration of these factors to future work. The present study focuses on a systematic noise-level sweep with a fixed, well-characterized architecture and training configuration optimized for reliable grokking behavior.

2.3.4 *Control experiments*

Baseline (no noise): Pure memorization behavior.

2.3.5 *Expected outcomes and hypotheses*

We hypothesized that, [1] The Grokking Index would peak at intermediate noise levels (5-10%). [2] Generalization delay would increase monotonically with noise level and [3] No grokking would be exhibited at some point after 10% noise.

2.3.6 *Success criteria*

Our objective was to observe grokking in all runs with $0\% \leq$ noise level $\leq 15\%$, achieving final test accuracy within 5% of training accuracy despite label noise.

2.3.7 *Failure modes*

We monitored for training instability at high noise levels and convergence failure where neither memorization nor generalization were achieved.

3. Results

3.1 Grokking behavior under label noise

The experimental investigation examined 155 separate neural networks runs across 31 different noise levels from 0% to 15% in 0.5% increments, with 5 independent runs per level. All runs exhibited grokking, defined as achieving training accuracy $\geq 95\%$, followed by test accuracy reaching within 5% of training accuracy after an extended period during which test accuracy remained significantly below training accuracy. Table 11 shows the effect of adding different amounts of training label noise on the grokking index, loss index, and the total epochs it took for grokking to occur.

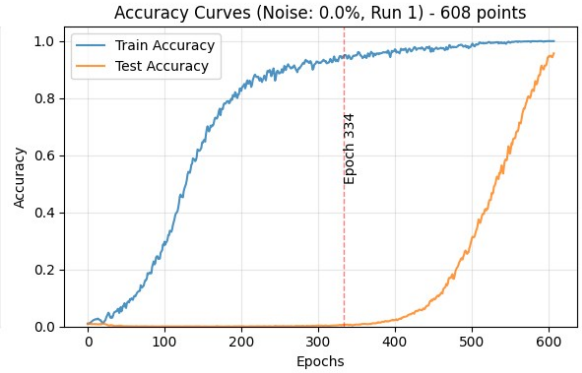

Figure 3. Sample run with 0% noise, with the training accuracy reaching 95% by epoch 334, and the test data catching up within 5% of the training accuracy in 608 epochs

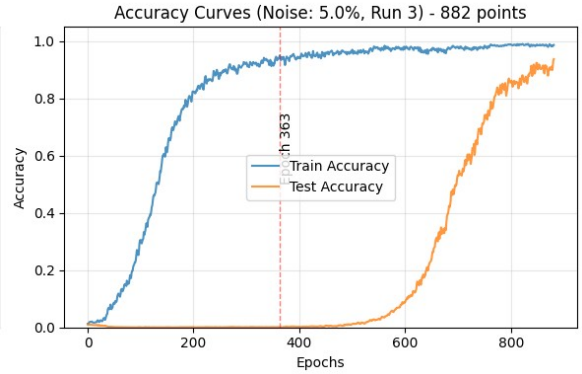

Figure 4. Sample run with 5% noise, with the training accuracy reaching 95% by epoch 363, and the test data catching up within 5% of the training accuracy in 882 epochs

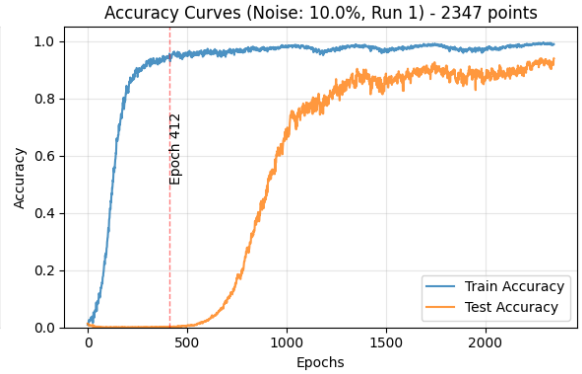

Figure 5. Sample run with 10% noise, with the training accuracy reaching 95% by epoch 412, and the test data catching up within 5% of the training accuracy in 2347 epochs

Table 11. Grokking with Label Noise Statistics. Shows Grokking Index, Loss Index, and Total Epochs for each 0.5% increase in noise percentage in training data from 0% to 15% along with their standard deviations, showing the spread of results between the 5 runs for each level.

Noise %	Grokking Index	Loss Index	Total Epochs
0	170.19±34.24	913.90 ± 294.81	588± 48
0.5	189.01± 50.66	1045.41 ± 373.80	617± 50
1	198.15± 29.10	1114.20 ± 251.19	646± 41
1.5	177.26± 21.56	913.73 ± 151.22	639± 39
2	201.8± 69.84	1127.36 ± 507.08	661± 95
2.5	232.35± 38.58	1285.69 ± 317.55	745± 44
3	249.64± 37.16	1421.78 ± 226.80	736± 67
3.5	257.73± 59.75	1328.94 ± 337.55	777± 101
4	253.91± 57.99	1403.54 ± 316.56	784± 108
4.5	297.32± 41.83	1802.02 ± 337.18	862± 48
5	337.8± 32.61	1899.83 ± 356.65	949± 85
5.5	400.14± 48.57	2282.12 ± 483.40	1109± 58
6	351.45± 47.64	2066.22 ± 442.17	974± 62
6.5	358.01± 75.36	1955.36 ± 430.64	1161± 307
7	376.95± 37.40	2033.44 ± 80.24	1264± 287
7.5	438.95± 88.61	2399.99 ± 555.14	1378± 249
8	470.5± 63.12	2649.34 ± 602.41	1686± 593
8.5	569.51± 243.53	3025.00 ± 1044.65	2331± 1566
9	686.94± 206.61	3239.45 ± 473.37	3346± 2093
9.5	908.99± 230.88	4111.10 ± 857.10	5378± 1768
10	960.46± 177.88	4468.17 ± 417.97	5095± 1798
10.5	1044.54± 134.71	4873.89 ± 962.07	5880± 429
11	1071.76± 82.06	5493.17 ± 322.16	4665± 703
11.5	1220.79± 218.02	5810.94 ± 481.71	5996± 2098
12	1431.08± 513.17	6194.69 ± 1589.92	6781± 2851
12.5	1398.98± 223.33	6165.18 ± 824.71	6901± 1310
13	1830.83± 374.28	7986.39 ± 1270.15	8756± 2042
13.5	2790.53± 1762.59	11337.80 ± 6155.91	17474 ± 1822
14	1860.04± 288.43	8246.71 ± 1221.41	8351± 1495
14.5	2544.45± 809.39	10689.88 ± 2701.40	11533± 4566
15	2751.87±304.17	11354.56 ±1468.15	12509± 1429

There were 3,764 training samples for the model, meaning that there are 19 corrupted sample for every 0.5% of the data (0.5% of 3764 = 18.82, rounded to 19). Every 1% is hence equivalent to 38 corrupted samples. The general relationship between noise and all three metrics exponentially increased, meaning that the amount of time to grok and the gap between memorization and performance scaled rapidly with each small increase in corrupted labels.

3.2 Relationship, variability, and local fluctuations

While the aggregate data supported exponential scaling, we noted local non-monotonicity in individual measurements. For instance, the Grokking Index at 1.5% noise (177.26±21.56) was lower than at 1% noise (198.15±29.10), and exhibited similar behavior at 6% versus 5.5% noise levels. These fluctuations reflect the stochastic nature of neural network training and the relatively small sample size (n=5) per noise level. Importantly, standard deviations increased significantly at higher noise levels— most notably at 13.5% (2790.53±1762.59)— indicating greater run-to-run variability as label corruption increased. Despite these local variations, the overall exponential trend remained statistically robust across the full range, as evidenced by the fitted models and aggregate statistics. Statistical validation using Levene's test confirmed heteroscedastic variance across noise levels (Levene statistic = 5.509, $p < 0.001$), with variability increasing systematically at higher noise. The data were well-fit by an exponential model with the following parameters and 95% confidence intervals:

$$\text{Grokking Index} = 97.5 \times e^{(0.224 \times \text{noise\%})} \quad \text{[b: 95\% CI: 0.199, 0.250; } R^2\text{=0.951]}$$
$$\text{Loss Index} = 642.4 \times e^{(0.193 \times \text{noise\%})} \quad \text{[b: 95\% CI: 0.175, 0.211; } R^2\text{=0.962]}$$
$$\text{Total Epochs} = 411.6 \times e^{(0.235 \times \text{noise\%})} \quad \text{[b: 95\% CI: 0.184, 0.286; } R^2\text{=0.855]}$$

All confidence intervals exclude zero, confirming statistically significant exponential scaling despite local fluctuations. The high R^2 values indicate the strong explanatory power of the exponential model across the noise range tested.

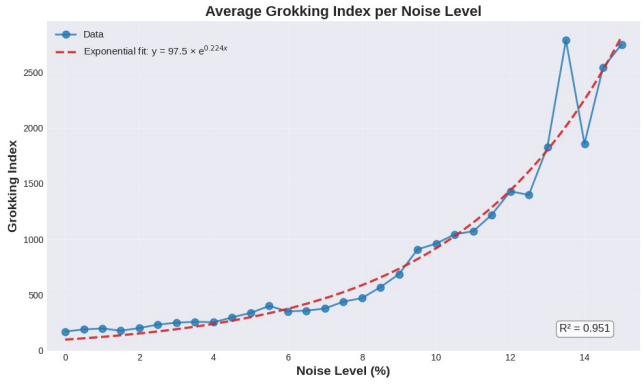

Figure 6. Grokking Index graphed over different levels of training data label noise

The grokking index versus percentage noise graph (Figure 6) showed an exponential relationship between the percentage of label noise added to training set and the total gap and length of the memorization and grokking period.

3.2.1 *Loss index results*

The loss index versus percentage noise (Figure 7) also showed an exponential relationship between the percentage of label noise added to the training set and the total gap and length of the memorization and grokking period.

3.2.2 *Total epochs results*

Similarly, the epoch versus percentage noise graph (Figure 8) showed an exponential relationship between the percentage of label noise added to the training set and the total time it took to grok.

Figure 7. Loss Index graphed over different levels of training data label noise

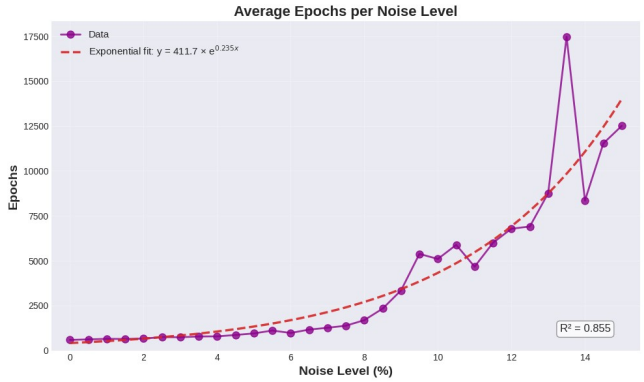

Figure 8. The total epochs for grokking to occur graphed over different levels of training data label noise

3.2.3 *All metrics relationship*

When the three main metrics were graphed superimposed on each other, (Figure 9) they all followed a similar pattern, especially the Loss Index and Grokking graphs. The three curves signify the same phenomenon, but each offers a different view.

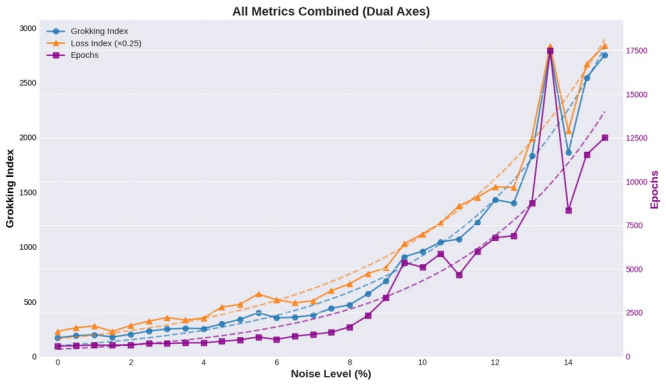

Figure 9. Grokking Index, 0.25 * Loss Index, and Total Epochs graphed over different levels of noise

3.2.4 *Robustness validation*

We performed two additional validation experiments examining robustness to train-test partition choice and hyperparameter configuration.

3.2.5 *Split validation across independent partitions*

We tested whether the exponential relationship depended on our specific train-test split by creating three independent random partitions of the 9,409 modular addition pairs. Visual inspection of sample examples (see section 7, Supplementary materials) confirmed that these were distinct partitions rather than mere re-initializations of the same split.

Table 12. Split validation across independent partitions

Metric	Noise %	Split 1	Split 2	Split 3	Between-Split SD	Within-Split SD	Ratio
Grokking Index	0	134.72±11.11	155.62±36.23	134.03±28.44	11.94	25.26	0.47
Grokking Index	5	316.09±53.72	304.13±27.25	281.99±32.42	14.30	37.80	0.38
Grokking Index	10	733.49±222.23	931.76±132.80	717.49±148.95	99.64	167.99	0.59
Grokking Index	15	4075.34±1508.25	3017.48±737.40	2815.95±693.33	572.02	979.66	0.58
Loss Index	0	672.16±44.69	826.64±229.76	713.65±229.01	66.83	167.82	0.40
Loss Index	5	1833.57±430.93	1648.28±149.77	1554.09±219.55	118.75	266.75	0.45
Loss Index	10	3476.82±914.82	4283.83±357.15	3549.46±434.44	371.75	568.80	0.65
Loss Index	15	16263.36±5279.71	12533.16±2650.44	11691.81±2681.37	2069.09	3536.92	0.58
Total Epochs	0	550.8±11.1	577.6±46.4	537.0±44.0	16.98	33.83	0.50
Total Epochs	5	901.8±88.2	902.8±81.7	843.2±42.3	28.33	70.73	0.40
Total Epochs	10	3880.6±1578.6	5139.8±1306.6	3532.2±1402.7	705.02	1429.30	0.49
Total Epochs	15	23022.0±13946.3	14071.2±4643.4	13281.0±3776.0	4535.75	7455.00	0.61

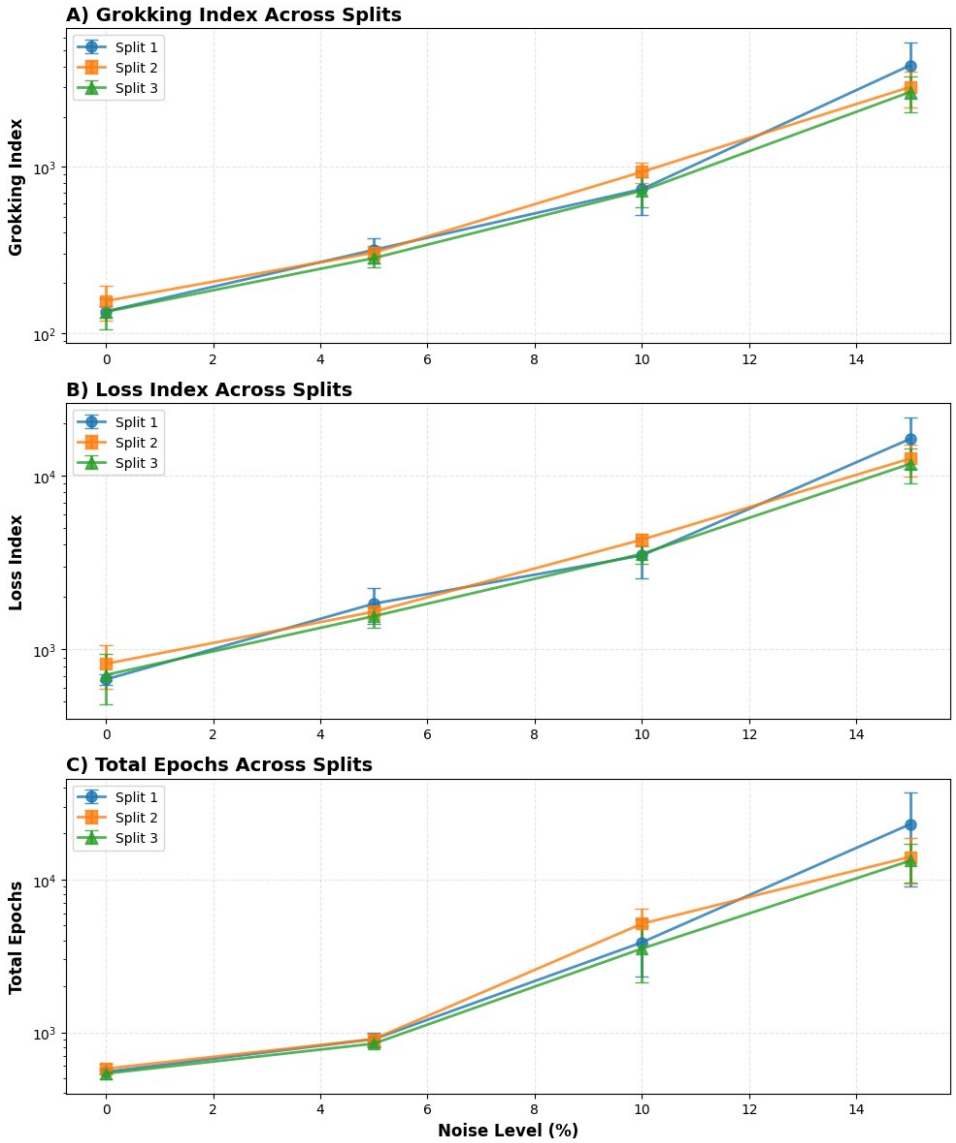

Figure 10. Grokking metrics across three independent random data partitions

Variance decomposition revealed that the between-split variability was moderate compared to within-split variability (Table 12). The ratio of between-split to within-split standard deviation averaged 0.51 for the Grokking Index, 0.52 for the Loss Index, and 0.50 for the Total Epochs across all noise levels. Ratios consistently < 1.0 indicated that choice of partition contributed ~ half as much variance as random initialization, confirming robustness to split selection.

All three splits exhibited exponential scaling of grokking metrics with noise level. At 0% noise, the three splits yielded Grokking Indices of 134.72±11.11, 155.62±36.23, and 134.03±28.44. At 15% noise, despite substantial absolute variation (4075.34±1508.25,

3017.48±737.40, and 2815.95±693.33), all splits maintained the exponential relationship. Figure 10 visualizes this consistency across splits for all three metrics, demonstrating that exponential scaling is not an artifact of our particular train-test partition but rather a robust property of the learning dynamics.

3.2.6 *Hyperparameter ablation*

To test whether exponential scaling depended on our specific optimization configuration, we systematically varied weight decay and learning rate across five configurations, testing each at noise levels of 0%, 5%, 10%, and 15% with 5 independent runs per condition (100 total runs).

Table 13. Hyperparameter ablation results

Metric	Configuration	Weight Decay	Learning Rate	0% Noise	5% Noise	10% Noise	15% Noise	Exponent (b)	95% CI	R^2
Grokking Index	Original	1.0	0.005	178.18±19.05	282.51±64.43	922.27±199.60	4237.10±2367.44	0.224	[0.199, 0.250]	0.951
Grokking Index	WD_0.5	0.5	0.005	773.31±109.38	1062.92±98.10	1690.62±107.48	4166.51±1468.94	0.121	[0.098, 0.144]	0.967
Grokking Index	WD_2.0	2.0	0.005	7.71±4.61	1.32±0.71	0.00±0.00	6957.62±905.45	—	—	—
Grokking Index	LR_1e-3	1.0	0.001	3821.51±779.95	5860.46±1723.29	28945.03±12129.81	44926.46±2714.94	—	—	—
Grokking Index	LR_1e-2	1.0	0.010	2.64±2.13	1.25±0.73	0.00±0.00	1421.55±2117.00	—	—	—
Loss Index	Original	1.0	0.005	1022.13±137.55	1585.12±475.54	4570.79±592.21	17657.64±8236.57	0.193	[0.175, 0.211]	0.962
Loss Index	WD_0.5	0.5	0.005	6142.39±1180.90	8391.88±785.00	12846.32±910.89	23775.64±5493.34	0.094	[0.081, 0.107]	0.988
Loss Index	WD_2.0	2.0	0.005	22.60±14.94	3.11±1.72	0.00±0.00	22857.63±2359.45	—	—	—
Loss Index	LR_1e-3	1.0	0.001	27173.74±6467.37	41494.34±13101.99	192930.57±114778.35	370006.60±70054.33	—	—	—
Loss Index	LR_1e-2	1.0	0.010	7.55±5.98	3.01±1.74	0.00±0.00	4819.32±7228.12	—	—	—
Total Epochs	Original	1.0	0.005	593.6±25.6	839.0±96.3	4560.8±1678.2	25455.4±20050.8	0.235	[0.184, 0.286]	0.855
Total Epochs	WD_0.5	0.5	0.005	1248.8±86.0	1699.8±192.2	2982.4±333.6	19349.4±15622.5	0.176	[0.139, 0.213]	0.922
Total Epochs	WD_2.0	2.0	0.005	359.2±22.8	433.2±47.0	643.0±102.6	50000.0±0.0	—	—	—
Total Epochs	LR_1e-3	1.0	0.001	5092.8±904.6	8803.8±1990.1	50000.0±0.0	50000.0±0.0	—	—	—
Total Epochs	LR_1e-2	1.0	0.010	389.0±59.2	504.0±42.7	715.2±77.8	14681.2±18342.0	—	—	—

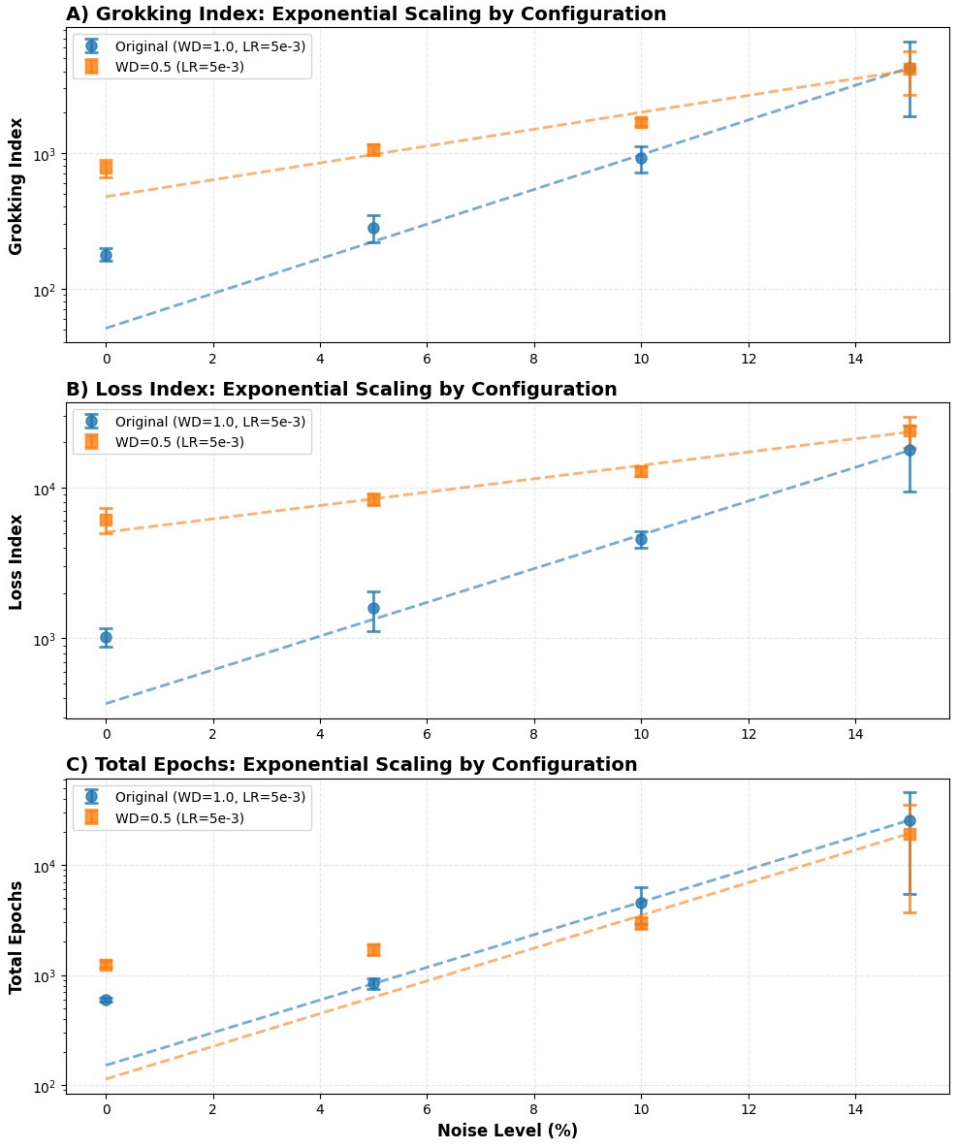

Figure 11. Comparison of grokking metrics under different weight decay settings

Configurations showing reliable grokking behavior (Original, WD_0.5) both exhibited exponential scaling relationships with high goodness-of-fit (Table 13). For the Original configuration, we observed: GI = 178.18±19.05 × $e^{(0.22\times noise\%)}$, R^2=0.951. The WD_0.5 configuration showed: GI = 773.31±109.38 × $e^{(0.121\times noise\%)}$, R^2=0.967. Critically, while absolute magnitudes varied substantially—with WD_0.5 showing 4.3× higher GI at 0% noise—both configurations exhibited positive exponential scaling exponents.

Three configurations (WD_2.0, LR_1x10^{-3}, LR_1x10^{-2}) showed extreme behavior: WD_2.0

achieved near-instant convergence at low noise (GI=7.71 at 0%) but failed to converge at 15% noise within 50,000 epochs; LR_1x10^{-3} showed significantly slow grokking (GI=44,926 at 15%); LR_1x10^{-2} exhibited non-monotonic behavior with near-instant convergence at low noise but inconsistent patterns at higher noise. These represent boundary cases where the optimization regime is poorly matched to the task, demonstrating limits to the robustness of grokking dynamics.

The scaling exponent b for successful configurations ranged from 0.094 (WD_0.5) to 0.235 (Original for Total Epochs), indicating different rates of noise-induced degradation (Figure 11). However, all stable configurations exhibited statistically significant positive exponential relationships (95% confidence intervals exclude zero), confirming that qualitative exponential scaling was robust to hyperparameter variations even as absolute magnitudes depended on the optimization regime.

These ablation studies demonstrated that while absolute grokking dynamics were sensitive to hyperparameter choices—as expected from optimization theory—the exponential relationship between label noise and grokking delay was a robust phenomenon across tested configurations that successfully induced grokking.

4. Discussion

Our investigation into the effects of label noise on neural network grokking dynamics revealed a striking exponential relationship between noise levels and both the severity and duration of delayed generalization. Our results contradicted all three of our initial hypotheses. Rather than finding an optimal noise level at 5-10%, we observed monotonic exponential degradation across the entire range. Contrary to our expectation that grokking might fail beyond 10% noise, all networks achieved grokking even at 15% noise, albeit with exponentially increased computational cost. These unexpected findings highlight the fragility of grokking dynamics and suggest that even minimal label corruption fundamentally alters the learning process.

4.1 Mechanistic interpretation

The exponential scaling of grokking metrics with noise levels suggests that label noise fundamentally disrupts the formation of generalizing circuits. Our results support a modified version of the circuit competition hypothesis, but with an important distinction: rather than noise preventing memorization circuits from forming, it appears to create inconsistent memorization targets that continuously interfere with the discovery of generalizing solutions.

Critically, our findings demonstrate that even minimal amounts of noise can be detrimental to grokking dynamics. With only 0.5% label corruption (approximately 19 mislabeled samples out of 3,764), we observed measurable delays in generalization. This challenges the intuitive notion that small amounts of noise might act as beneficial regularization. Instead, our results suggest that for algorithmic tasks, any deviation from the true underlying pattern creates additional computational burden that compounds exponentially. This finding

contrasts sharply with studies suggesting that certain forms of "internal noise" or stochasticity during training can improve generalization—in our experiments, such noise appears uniformly harmful to the grokking process.

The model must simultaneously contend with two challenges: learning the true underlying modular arithmetic operation and filtering out the corrupted labels. This dual burden manifests in the observed exponential delays, as the network requires increasingly more evidence to distinguish signal from noise. The high correlation between our Grokking Index and Loss index (Figure 9) further supports this interpretation, suggesting that the confidence gap directly reflects the model's uncertainty about which patterns to trust.

4.2 Scope and generalizability

Our findings are conditional on the specific experimental configuration employed; a weight decay of 1.0, a learning rate of 5×10^{-3}, a 40/60 train-test split, and uniform random label noise. While we observe robust exponential scaling within this setup, the absolute dynamics and scaling exponents may differ under alternative hyperparameter regimes. The interaction between label noise and other regularization mechanisms (weight decay, learning rate scheduling) means that our causal claims apply specifically to this configuration. We maintained a clean test set following standard practice in noisy-label learning to ensure accurate generalization measurement; applying noise to the test set would conflate learning difficulty with evaluation uncertainty. Our 0-15% noise range captures low-to-moderate corruption and reveals monotonic degradation, but we cannot rule out regime changes at higher noise levels or under different data splits. These results are specific to modular addition, ML architecture and require validation on other algorithmic tasks before broader generalization.

4.3 Reconciling with previous literature

Our findings reveal a fundamental distinction between standard generalization and grokking with respect to noise. While controlled noise demonstrably helps standard generalization—through dropout, label smoothing, and noisy student training—it uniformly harms grokking. This contradiction suggests grokking operates through fundamentally different mechanisms than standard generalization. Previous studies examining beneficial noise effects typically focused on preventing immediate overfitting in continuous optimization landscapes. In contrast, the modular arithmetic task presents a discrete, highly structured problem where even small amounts of label corruption directly violate the mathematical relationships the model must learn. The absence of any "sweet spot" for noise-induced grokking in our experiments (0-15% range) suggests that for algorithmic tasks requiring precise pattern recognition, label noise uniformly degrades learning dynamics. This aligns with theoretical work on the conditions necessary for grokking, where grokking requires a delicate balance between competing loss landscape features—a balance that label noise appears to systematically disrupt.

4.4 Why dropout helps but label noise hurts: a mechanistic explanation

The distinction between beneficial noise (like dropout) and detrimental noise (label corruption) in grokking contexts can be

understood through their fundamentally different mechanisms of action. Dropout operates as a transient perturbation that affects different neurons on each forward pass, forcing the network to develop redundant representations that don't rely on any single pathway. This stochastic masking prevents co-adaptation between neurons and encourages robust feature learning. Crucially, dropout preserves the true input-output mapping—the correct answer for any given input remains constant throughout training. In contrast, label noise creates persistent contradictions in the training data. When 5% of labels are corrupted, the network repeatedly encounters cases where identical inputs (e.g., 23 + 45) map to different outputs across epochs. This forces the network to memorize exceptions rather than learn the underlying modular arithmetic rule. While dropout says, "find multiple ways to compute the right answer," label noise says, "learn that sometimes 23 + 45 = 68 and sometimes 23 + 45 = 17." For grokking specifically, this distinction becomes critical. Grokking requires the network to discover a clean algorithmic solution—essentially reverse-engineering the modular addition operation. This discovery process depends on recognizing consistent patterns across the dataset. Dropout does not interfere with this pattern recognition; it merely ensures the patterns are learned robustly. Label noise, however, obscures the very patterns that the network needs to discover. Each corrupted label represents a false counterexample to the true algorithm, exponentially increasing the evidence required to confidently extract the underlying rule. This explains our exponential scaling law: each additional corrupted label does not just add one more exception to memorize but multiplicatively increases the hypothesis space the network must search. With 0% noise, there is exactly one consistent algorithm (modular addition). With 5% noise, the network must consider significantly more hypotheses that could explain the contradictory data, thereby delaying the generalization.

4.5 Mechanistic hypothesis for exponential scaling

We propose that the exponential relationship arises from the combinatorial explosion of hypothesis space under label noise. Consider that the network must identify the true modular addition rule from contradictory evidence. With clean data, there is exactly one consistent hypothesis: f(a,b) = (a+b) mod 97.

With k corrupted labels, the network encounters k contradictions that must be explained. Each contradiction could be an error to be ignored (the correct approach), an exception to be memorized, or evidence for a different rule entirely.

The network cannot *a priori* distinguish corrupted from clean labels. Therefore, it must consider hypotheses that explain various subsets of the data. With k corruptions, there are 2^k possible subsets that might represent the "true" pattern, leading to exponential growth in hypothesis space. This combinatorial explosion manifests in our data: the scaling exponents (0.224 for Grokking Index, 0.235 for Total Epochs) suggest that each 1% increase in noise (~38 samples) multiplies the search space by approximately 1.2x. While we cannot prove this mechanism without analyzing internal representations, it provides a testable

framework: tasks with larger hypothesis spaces should show steeper exponential scaling.

4.6 Implications for understanding grokking

Our results provide empirical evidence that grokking may not merely a consequence of extended training but requires specific conditions in the loss landscape that guide the transition from memorization to generalization. The exponential scaling laws suggest that each increment of noise multiplicatively increases the difficulty of discovering generalizing solutions.

This has important theoretical implications for understanding the grokking phenomenon. Rather than viewing grokking as a robust emergent property of overparameterized networks, our findings suggest it is a fragile phenomenon that depends critically on the consistency of training signals. The fact that all networks eventually achieved grokking despite noise levels up to 15% demonstrates the remarkable ability of neural networks to eventually extract signal from noise, but the exponential cost in training time raises questions about the practical viability of this learning regime.

4.7 Practical considerations

For practitioners working with potentially noisy datasets, our findings sound a cautionary note about expecting delayed generalization to rescue poor early training performance. The exponential scaling means that even modest label noise (5%) can increase the time to generalization by nearly 6-fold, with 10% noise requiring almost 9 times longer to achieve comparable generalization. Due to the exponential nature of this relationship, each 0.5% increment in noise does not just add a fixed delay—it multiplicatively increases the computational burden. This has profound implications for training budgets: what might be a feasible 600-epoch training run at 0% noise may translate into a 12,500-epoch marathon at 15% noise, representing a 20-fold increase in computational resources required.

These results suggest that for algorithmic tasks or problems with known ground-truth structure, investing in data quality and label verification may be more efficient than relying on extended training to overcome noise through grokking. However, our work also demonstrates that neural networks can, given sufficient time, extract meaningful patterns even from significantly corrupted training data—a finding that may inform strategies for learning from weakly supervised or crowd-sourced labels.

4.8 Limitations and considerations

Several limitations of our study should be acknowledged. First, we examined only uniform random label noise, whereas real-world label corruption often exhibits structure or bias. Second, our investigation focused solely on modular addition with a specific architecture and training regime. The exponential relationships we observed may not generalize to other algorithmic tasks, different architectures, or alternative optimization procedures.

Additionally, our noise range (0-15%) was chosen to maintain eventual convergence, but higher noise levels (>15%) might reveal different dynamics, phase transitions, or even non-monotonic relationships where noise could potentially benefit grokking. We cannot rule out

the existence of beneficial noise regimes at corruption levels beyond our tested range, particularly under different train-test split ratios or alternative hyperparameter configurations.

Our sample size of five runs per noise level, while sufficient to detect strong trends and establish statistical significance, limits statistical power for detecting subtle effects and contributes to the local non-monotonicities observed in Table 11. Larger sample sizes would provide more stable estimates and tighter confidence intervals, particularly at higher noise levels where variance increases substantially.

While we validated robustness across three independent splits and five hyperparameter configurations (demonstrating persistent exponential scaling), our findings are still specific to modular addition with MLPs. Future work should extend to other algorithmic tasks and architectural choices to establish (or not) full generality of these scaling laws.

4.9 Perspectives

Several natural extensions of this work would strengthen and broaden our findings. Testing other algorithmic operations such as modular multiplication, modular subtraction, or permutation group tasks would establish whether the noise-grokking relationship generalizes beyond modular addition. Exploring different architectural families (transformers, convolutional networks) would reveal whether exponential scaling is architecture-independent or specific to MLPs.

Our findings open several avenues for future research. Investigating whether the exponential scaling laws hold across different algorithmic tasks would establish the generality of our results. Exploring structured noise patterns, such as class-conditional corruption or adversarial label flips, could reveal whether certain types of noise are less detrimental to grokking dynamics.

Theoretical work is needed to explain why the relationship is specifically exponential rather than (say) polynomial or logarithmic. This might involve analyzing the loss landscape geometry under label noise or developing mean-field theories that account for noise-induced perturbations in the learning dynamics.

From a practical standpoint, developing training strategies that mitigate the effects of label noise on grokking could be valuable. Curriculum learning approaches that gradually introduce noisy examples, or adaptive weighting schemes that identify and down-weight corrupted samples, might preserve some of the benefits of delayed generalization while reducing the exponential cost of noise.

5. Conclusions

This study provides the first systematic characterization of how label noise affects neural network grokking dynamics. The exponential deterioration in grokking behavior with increasing noise levels reveals the fragility of this phenomenon and challenges assumptions about its robustness. Our key finding—that the relationship follows exponential scaling has critical implications for both theory and practice.

The exponential nature of this relationship means that small increments in label noise have disproportionately large effects on training dynamics. Even a seemingly negligible 0.5% increase in noise can translate to hundreds of additional epochs required for generalization, while moving from 10% to 15% noise more than doubles the time to grok (from ~5,000 to ~12,500 epochs). This exponential scaling suggests that there is no "safe" level of label noise for algorithmic tasks—every corrupted label compounds the difficulty of discovering generalizing solutions in a multiplicative fashion.

While our results demonstrate that neural networks can eventually overcome significant label corruption through extended training, the exponential cost suggests that for practical applications, ensuring data quality remains paramount. The absence of any beneficial effect from small amounts of noise contradicts the notion that controlled corruption might serve as a useful regularizer for inducing grokking. Instead, our findings indicate that label noise uniformly impedes the delicate transition from memorization to generalization that characterizes the grokking phenomenon.

These findings contribute to our understanding of the fundamental tension between memorization and generalization in deep learning and highlight the need for continued research into the conditions that enable neural networks to discover generalizing solutions from imperfect training data. Future work should investigate whether alternative training strategies or architectural innovations can mitigate the exponential cost of learning from noisy labels while preserving the beneficial aspects of delayed generalization.

Acknowledgement

I would like to express my gratitude to Shaddin Dughmi of the University of Southern California, Department of Computer Science, for his mentorship and guidance throughout this project.

6. Data and code availability

All experimental code, training scripts, and complete dataset splits are publicly available at https://github.com/havishs99/Investigating-the-Impact-of-Low-to-Moderate-Level-Label-Noise-on-Neural-Network-Grokking-Dynamics

Specifically, the repository includes, [1] Main experiment results (155 training runs with epoch-by-epoch metrics for 31 noise levels from 0% to 15%), [2] Split validation results (60 training runs across 3 splits × 4 noise levels × 5 runs), [3] Hyperparameter ablation results (100 training runs across 5 configurations × 4 noise levels × 5 runs), [4] Analysis scripts for computing Grokking Index, Loss Index, exponential fits, and variance decomposition, and [5] Figure generation code for all visualizations.

7. Supplementary materials

Split Fingerprints (Verification of Independence)

To verify that our three splits represent genuinely independent partitions rather than mere re-initializations of the same split, we document the first 10 training and test examples from each split.

Split 1 - First 10 *Training* Examples: (26,30)→56, (19,37)→56, (65,46)→14, (36,30)→66, (67,57)→27, (15,62)→77, (85,32)→20, (87,82)→72, (34,12)→46, (2,42)→44
Split 1 - First 10 *Test* Examples: (79,55)→37, (18,13)→31, (90,76)→69, (70,27)→0, (63,58)→24, (33,19)→52, (66,94)→63, (2,67)→69, (70,14)→84, (24,41)→65

Split 2 - First 10 *Training* Examples: (13,7)→20, (83,52)→38, (83,50)→36, (92,31)→26, (11,48)→59, (43,18)→61, (5,87)→92, (44,60)→7, (61,13)→74, (94,32)→29
Split 2 - First 10 *Test* Examples: (29,76)→8, (92,54)→49, (44,62)→9, (72,19)→91, (15,53)→68, (25,93)→21, (54,73)→30, (39,26)→65, (44,95)→42, (33,5)→38

Split 3 - First 10 *Training* Examples: (12,9)→21, (48,6)→54, (75,77)→55, (52,10)→62, (76,11)→87, (48,83)→34, (91,38)→32, (55,52)→10, (14,73)→87, (77,21)→1
Split 3 - First 10 *Test* Examples: (35,94)→32, (58,11)→69, (91,37)→31, (66,66)→35, (39,5)→44, (81,28)→12, (8,90)→1, (10,10)→20, (46,34)→80, (7,17)→24

8. References

1. Power, A., Burda, Y., Edwards, H., Babuschkin, I., Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. arXiv preprint arXiv:2201.02177. https://arxiv.org/abs/2201.02177

2. Nanda, N., Chan, L., Lieberum, T., Smith, J., Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. International Conference on Learning Representations. https://openreview.net/forum?id=9XFSbDPmdW

3. Liu, Z., Kitouni, O., Nolte, N., Michaud, E. J., Tegmark, M., Williams, M. (2023). Towards understanding grokking: An effective theory of representation learning. Advances in Neural Information Processing Systems, 36. https://proceedings.neurips.cc/paper_files/paper/2022/hash/dfc310e81992d2e4cedc09ac47eff13e-Abstract-Conference.html

4. Frénay, B., Verleysen, M. (2014). Classification in the presence of label noise: A survey. IEEE Transactions on Neural Networks and Learning Systems, 25, 845-869. https://doi.org/10.1109/TNNLS.2013.2292894

5. Song, H., Kim, M., Park, D., Shin, Y., Lee, J. G. (2022). Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems, 33, 5737-5754. https://doi.org/10.1109/TNNLS.2022.3152527

6. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. International Conference on Learning Representations. https://openreview.net/forum?id=Sy8gdB9xx

7. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., et al. (2017). A closer look at memorization in deep networks. International Conference on Machine Learning, 233-242. https://proceedings.mlr.press/v70/arpit17a.html

8. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15, 1929-1958. http://jmlr.org/papers/v15/srivastava14a.html

9. Müller, R., Kornblith, S., Hinton, G. (2019). When does label smoothing help? Advances in Neural Information Processing Systems, 32. https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html

10. Cubuk, E. D., Zoph, B., Shlens, J., Le, Q. V. (2020). RandAugment: Practical automated data augmentation with a reduced search space. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 702-703. https://doi.org/10.1109/CVPR42600.2020.01070

11. Xie, Q., Luong, M. T., Hovy, E., Le, Q. V. (2020). Self-training with noisy student improves ImageNet classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10687-10698. https://doi.org/10.1109/CVPR42600.2020.01070

12. Liu, Z., Kitouni, O., Nolte, N., Michaud, E. J., Tegmark, M., Williams, M. (2022). Omnigrok: Grokking beyond algorithmic data. arXiv preprint arXiv:2210.01117. https://arxiv.org/abs/2210.01117

13. Varma, V., Shah, R., Kenton, Z., Křmela, J., Kumar, R. (2023). Explaining grokking through circuit efficiency. arXiv preprint arXiv:2309.02390. https://arxiv.org/abs/2309.02390

14. Kumar, K., Verma, A., Srinivasan, A., Bengio, Y. (2024). Grokking as the transition from lazy to rich training dynamics. International Conference on Learning Representations. https://openreview.net/forum?id=vTBIJmMgzm

15. Lyu, K., Jin, Z., Li, Z., Du, S. S., Lee, J. D., Hu, W. (2024). Dichotomy of early and late phase implicit biases can provably induce grokking. International Conference on Learning Representations. https://openreview.net/forum?id=X0UD5XZFY1

16. Natarajan, N., Dhillon, I. S., Ravikumar, P. K., Tewari, A. (2013). Learning with noisy labels. Advances in Neural Information Processing Systems, 26. https://proceedings.neurips.cc/paper/2013/hash/3871bd64012152bfb53fdf04b401193f-Abstract.html

17. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1944-1952. https://doi.org/10.1109/CVPR.2017.240

18. Miller, J., Rae, J., Borgeaud, S., Michaud, E., Hill, F. (2023). Grokking in linear estimators – a solvable model that groks without understanding. arXiv preprint arXiv:2310.16441. https://arxiv.org/abs/2310.16441

19. Millidge, B., Kazuki, O., McLeish, T. (2023). Grokking as compression: A statistical mechanics perspective. arXiv preprint arXiv:2310.05918. https://arxiv.org/abs/2310.05918

20. Thilak, V., Littwin, E., Zhai, S., Saremi, O., Paiss, R., Susskind, J. (2022). The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. arXiv preprint arXiv:2206.04817. https://arxiv.org/abs/2206.04817

21. Notsawo, F., Rousseau, R., Roueff, F., Vaiter, S. (2024). Predicting grokking long before it happens: A look into the loss landscape of models which grok. International Conference on Machine Learning. https://proceedings.mlr.press/v235/notsawo24a.html

22. Humayun, A. I., Balestriero, R., Baraniuk, R. (2024). Deep Networks Always Grok and Here is Why. International Conference on Machine Learning. https://arxiv.org/abs/2402.15555

23. Chen, A., Shwartz-Ziv, R., Cho, K., Leavitt, M. L., Saphra, N. (2024). Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs. International Conference on Learning Representations. https://arxiv.org/abs/2309.07311