Quantifying the Memorization-to-Generalization Transition: Scaling Laws and Phase Structure in Grokking

Anonymous Authors[1]

Abstract

Neural networks trained past memorization frequently undergo a delayed transition to generalization, a phenomenon known as grokking. Despite theoretical progress on *why* this transition occurs, the quantitative structure of *when* it occurs in hyperparameter space remains uncharacterized. We map the memorization-to-generalization boundary across 384 configurations of two-hidden-layer MLPs on modular arithmetic, fitting a power-law scaling relation for generalization onset time: $T_{\text{grok}} \propto H^{-0.27} D^{-2.04} \eta^{-0.50} \lambda^{-0.64}$ ($R^2 = 0.732$; 0.821 with interactions). The exponent hierarchy reveals that data complexity ($D^{-2.04}$) is the dominant driver of regime transition, not model capacity ($H^{-0.27}$): doubling data accelerates generalization by $\sim 4\times$, while doubling width yields only $\sim 1.2\times$. A sharp phase boundary at weight decay $\lambda \gtrsim 1.0$ separates grokking from non-grokking configurations, and weight norm trajectories show monotonic compression during the transition, consistent with implicit regularization selecting low-complexity solutions. These results provide a quantitative foundation for predicting and controlling regime transitions in overparameterized networks.

1. Introduction

When do overparameterized neural networks transition from memorizing their training data to genuinely generalizing? This question sits at the center of modern deep learning theory. The double descent phenomenon (Belkin et al., 2019; Nakkiran et al., 2020) shows that generalization can improve *beyond* the interpolation threshold, and the lazy-to-rich transition (Woodworth et al., 2020; Chizat et al., 2019) identifies feature learning as the mechanism that separates kernel-regime memorization from structured generalization. Yet these frameworks characterize the transition qualitatively; they do not predict *when*, in training time, a given configuration will cross from one regime to the other.

Grokking (Power et al., 2022) provides a clean experimental window into this question. Networks trained on modular arithmetic first memorize the training set perfectly, then, after continued training with weight decay, abruptly generalize. The delay between memorization and generalization can span orders of magnitude depending on hyperparameters. Mechanistic studies have identified *what* happens during this transition: Fourier-feature circuits replace memorization lookup tables (Nanda et al., 2023), competing subnetworks resolve in favor of sparse generalizing solutions (Merrill et al., 2023), and weight norms compress toward low-complexity basins (Liu et al., 2023).

What is missing is a *quantitative scaling theory*: which hyperparameters govern the transition time, in what proportion, and does a phase boundary exist in hyperparameter space? We address this with a 384-configuration sweep, fitting a power-law scaling law for generalization onset time T_{grok}.

Our contributions:

1. A power-law scaling relation $T_{\text{grok}} \propto H^{-0.27} D^{-2.04} \eta^{-0.50} \lambda^{-0.64}$ ($R^2 = 0.821$ with interactions), revealing that data complexity dominates model capacity by an order of magnitude in exponent.

2. A sharp phase boundary at $\lambda \gtrsim 1.0$ in (η, λ)-space, below which generalization is structurally suppressed regardless of training duration.

3. Weight norm dynamics showing monotonic compression (median ratio $\|\theta(T_{\text{grok}})\| / \|\theta(T_{\text{mem}})\| = 0.42$) as a measurable proxy for regime state.

The central finding overturns a natural intuition: it is data coverage and regularization strength, not model capacity,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

that primarily control when networks escape the memorizing regime.

2. Experimental Setup

Tasks. We study addition mod 113 ($113^2 = 12{,}769$ examples) and division mod 97 ($97 \times 96 = 9{,}312$ examples), canonical grokking benchmarks (Power et al., 2022) where the ground-truth generalizing solution (discrete Fourier transform) is known (Nanda et al., 2023).

Architecture. A two-hidden-layer MLP with learned embeddings: $f(a, b) = W_3\,\sigma(W_2\,\sigma(W_1[e_a; e_b]))$, where $e_a, e_b \in \mathbb{R}^H$, σ is ReLU, and $H \in \{128, 256, 512\}$ (100K–960K parameters).

Hyperparameter sweep. We vary data fraction $D \in \{0.3, 0.5, 0.7, 0.97\}$, learning rate $\eta \in \{0.001, 0.003, 0.01, 0.03\}$, and weight decay $\lambda \in \{0.1, 0.3, 1.0, 3.0\}$, yielding $2 \times 3 \times 4^3 = 384$ configurations. All models use AdamW (β_1=0.9, β_2=0.98), full-batch training, up to 150K steps, each with a single random seed (median seed variance CV $\approx 8\%$; up to 18% near the phase boundary, quantified in Appendix A). Of 356 completed runs (28 diverged at high η, low λ), 297 (83.4%) grokked.

Regime operationalization. We define T_{mem} as the first step with training accuracy >99% and T_{grok} as the first step with test accuracy >95%. A configuration is *non-grokking* if $T_{\text{grok}} > 150{,}000$. This operationalization cleanly separates the memorization plateau from the generalization transition: the grokking gap $\Delta T = T_{\text{grok}} - T_{\text{mem}}$ spans from 100 to over 100,000 steps ($\sim$1,000$\times$ range).

3. Scaling Law for Generalization Onset

We fit a log-linear model over the 297 grokked runs:

$$\log T_{\text{grok}} = \alpha \log H + \beta \log D + \gamma \log \eta + \delta \log \lambda + c\,, \quad (1)$$

yielding the power-law form

$$T_{\text{grok}} \;\propto\; H^{-0.27}\, D^{-2.04}\, \eta^{-0.50}\, \lambda^{-0.64} \quad (2)$$

with $R^2 = 0.732$. All exponents are negative: increasing any hyperparameter reduces T_{grok}.

3.1. Exponent Hierarchy: Data Dominates Capacity

The exponent magnitudes in Table 1 establish a clear hierarchy among the drivers of generalization onset.

Data complexity ($D^{-2.04}$). Doubling the training fraction cuts T_{grok} by $2^{2.04} \approx 4.1\times$. This superlinear dependence suggests that additional data does more than provide redundant examples: it simultaneously strengthens the signal for the generalizing circuit and destabilizes the memorizing solution by reducing its effective capacity advantage. In the Fourier-feature picture of Nanda et al. (Nanda et al., 2023), each training pair constrains the phase of a Fourier component; at high D, the constraint set becomes overdetermined for the memorizing lookup table but remains consistent for the algebraic circuit.

Table 1. Fitted exponents with standard errors and Spearman correlations. Data fraction D has the steepest exponent and second-highest rank correlation; weight decay λ has the highest rank correlation despite a smaller exponent, reflecting its role in the phase boundary.

	H	D	η	λ
Exponent	-0.27	-2.04	-0.50	-0.64
	± 0.10	± 0.12	± 0.04	± 0.05
Spearman ρ	-0.08	-0.41	-0.28	-0.52

Implicit regularization ($\lambda^{-0.64}$). Weight decay is the second lever. Its Spearman correlation with T_{grok} ($\rho = -0.52$) is higher than that of any other hyperparameter, reflecting its dual role: λ both accelerates the transition *and* determines whether it occurs at all (Section 4). The $\lambda^{-0.64}$ exponent is consistent with implicit regularization theory (Lyu & Li, 2020): weight decay biases gradient descent toward low-norm solutions, and the rate of this bias scales sublinearly with λ.

Learning rate ($\eta^{-0.50}$). The $\eta^{-0.50}$ scaling is consistent with SGD convergence rates: faster optimization traverses the loss landscape more quickly, reducing the time to reach the generalizing basin. The interaction term $\log H \times \log \eta$ ($t = 6.2$) indicates wider models benefit more from higher learning rates, suggesting the optimization landscape becomes more navigable with increased capacity.

Model capacity ($H^{-0.27}$). Width has the weakest effect. Doubling H reduces T_{grok} by only $2^{0.27} \approx 1.2\times$. This contradicts the naive expectation that larger models should generalize faster and supports a view where the transition is governed by *optimization dynamics* (how quickly weight decay compresses the network) rather than *representational capacity* (how many parameters are available). The low Spearman correlation ($\rho = -0.08$) confirms that width's contribution is further muted by interactions with other hyperparameters.

3.2. Interaction Effects

Adding all six pairwise interactions plus a binary task indicator raises R^2 to 0.821 (adjusted $R^2 = 0.813$; leave-one-run-out $R^2 = 0.799$). Four interactions are significant (all

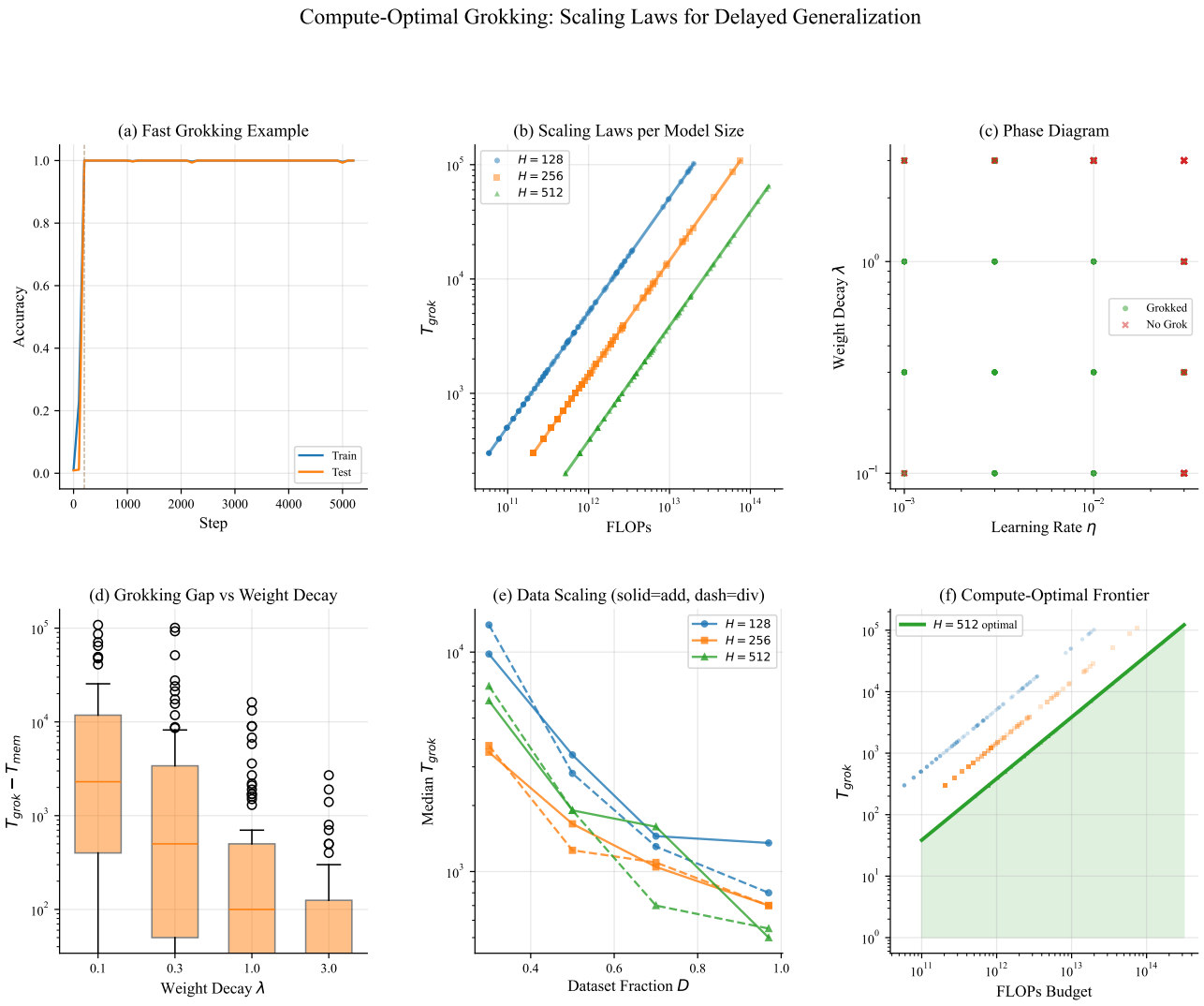

Figure 1. **Grokking dynamics across 356 runs, two tasks, and three model sizes.** (a) Representative training curves showing the memorization plateau followed by delayed generalization. (b) T_{grok} vs. FLOPs by model width: wider models grok faster in steps but at higher FLOP cost. (c) Phase diagram in (η, λ) space: grokking rate approaches 100% for $\lambda \geq 1.0$. (d) Grokking gap vs. λ by width. (e) Data scaling: median T_{grok} vs. D. (f) Compute-optimal frontier.

$|t| > 4$):

- $\log D \times \log \eta$: $+0.50$ (t=6.3). At high data fractions, faster learning rates accelerate grokking more.
- $\log H \times \log \eta$: $+0.35$ (t=6.2). Wider models benefit more from higher learning rates.
- $\log D \times \log \lambda$: $+0.38$ (t=5.3). Larger datasets partially substitute for strong regularization.
- $\log H \times \log \lambda$: -0.23 (t=-4.1). Wider models are less sensitive to weight decay.

The $D \times \lambda$ interaction ($+0.38$) is notable: when data is abundant ($D \to 1$), the marginal effect of weight decay is reduced. This supports a picture where data coverage and regularization are *partially substitutable* mechanisms for driving the memorization-to-generalization transition, both acting to destabilize the high-norm memorizing solution.

3.3. Robustness

The exponents are stable across tasks (addition: $D = -1.95$; division: $D = -2.12$; pooled: -2.04) and across generalization thresholds (Appendix A). A Weibull accelerated failure time model fit to all 356 runs (59 right-censored) achieves concordance index 0.71 with coefficient signs consistent with OLS. The Weibull shape parameter $k = 1.4$ implies a mildly increasing hazard: once a net-

work begins transitioning, it accelerates, consistent with the positive-feedback picture of norm compression (Section 5).

4. Phase Structure of the Grokking Boundary

The scaling law describes how T_{grok} varies *within* the grokking regime. Equally important is the boundary *between* regimes: which configurations grok at all?

Figure 2 (left) shows a sharp phase boundary in (η, λ)-space. At $\lambda \geq 1.0$, nearly all configurations grok regardless of learning rate. At $\lambda = 0.1$, fewer than 60% do, and grokking becomes sensitive to η. The boundary is not gradual: the transition from $<$60% to $>$95% grokking occurs within a factor of $3\times$ in λ.

Conjecture 1 (Phase boundary)**.** *There exists a critical regularization strength $\lambda^* \approx 1.0$ (relative to the AdamW scale used here) such that generalization onset is structurally suppressed for $\lambda < \lambda^*$. Below λ^*, the memorizing solution is a stable fixed point of the training dynamics; above λ^*, weight decay destabilizes the memorizing basin, and the network converges to a lower-norm generalizing solution.*

This conjecture is consistent with the Omnigrok framework of Liu et al. (Liu et al., 2023), where grokking occurs when the weight norm trajectory crosses a critical threshold at which the generalizing loss landscape basin becomes accessible. Our phase diagram provides the first quantitative characterization of where this threshold sits in hyperparameter space.

The grokking gap. The delay $\Delta T = T_{\text{grok}} - T_{\text{mem}}$ drops by more than an order of magnitude as λ increases from 0.1 to 3.0. This gap measures the time spent in the memorizing regime after the training objective is satisfied; it is the "wasted" computation from a generalization perspective. The scaling law predicts this gap, enabling practitioners to estimate whether a given configuration will exhibit delayed generalization or rapid transition.

5. Weight Norm Dynamics as a Regime Indicator

Existing accounts of grokking emphasize that norm compression drives the transition (Liu et al., 2023; Merrill et al., 2023). We quantify this directly.

From 20 configurations stratified by T_{grok} (fast $<$ 1K, medium 1K–10K, slow $>$ 10K; draws within each stratum), we track weight norms at 100-step intervals. Among the 14 of 20 runs with non-trivial grokking gap, the norm at generalization is lower than at memorization in all 14 cases. The median ratio $\|\theta(T_{\text{grok}})\|/\|\theta(T_{\text{mem}})\| = 0.42$ (IQR: 0.31–0.54). While the sample is modest, the monotonic compression pattern is consistent across all three strata (fast, medium, slow) and both tasks.

Conjecture 2 (Norm compression threshold)**.** *The memorization-to-generalization transition occurs when the weight norm ratio $\|\theta(t)\|/\|\theta(T_{\text{mem}})\|$ crosses a threshold r^* in the range 0.3–0.5. If r^* is approximately configuration-independent, then the time to reach it is governed by the scaling law* (2)*, with weight decay λ controlling the compression rate and data fraction D controlling how much compression is needed.*

Two observations support this conjecture. First, the norm ratio is monotonically decreasing in all 14 runs during the memorization phase: the network progressively compresses before the transition. Second, the absolute norm at T_{mem} does not predict T_{grok} (Spearman $\rho = 0.06$, $p = 0.83$), but the *rate* of norm decrease does: configurations with faster compression grok sooner. This is consistent with the view that what matters is not the starting point but the *trajectory* through weight space, governed by the interplay of λ (compression force) and D (landscape structure).

6. Related Work

Grokking. Power et al. (Power et al., 2022) discovered grokking in modular arithmetic. Nanda et al. (Nanda et al., 2023) reverse-engineered the Fourier circuits that form during the transition. Liu et al. (Liu et al., 2023) showed grokking extends beyond algorithmic data and proposed the norm-based LU mechanism. Barak et al. (Barak et al., 2022) demonstrated hidden progress in parity learning with SGD. These works characterize *what* happens; we characterize *when*, providing the first quantitative scaling law.

Implicit regularization and regime transitions. Lyu and Li (Lyu & Li, 2020) showed gradient descent converges to max-margin solutions for homogeneous networks. Woodworth et al. (Woodworth et al., 2020) identified the lazy-to-rich transition as a function of initialization scale. Our $\lambda^{-0.64}$ exponent quantifies how explicit regularization modulates this transition in practice.

Phase transitions in learning. Belkin et al. (Belkin et al., 2019) and Nakkiran et al. (Nakkiran et al., 2020) identified the interpolation threshold as a phase boundary. Our phase diagram (Figure 2) identifies an analogous boundary in (η, λ)-space governing delayed generalization.

Scaling laws. Kaplan et al. (Kaplan et al., 2020) and Hoffmann et al. (Hoffmann et al., 2022) established power-law scaling for language model loss. Bordelon et al. (Bordelon et al., 2024) derived scaling laws from a solvable

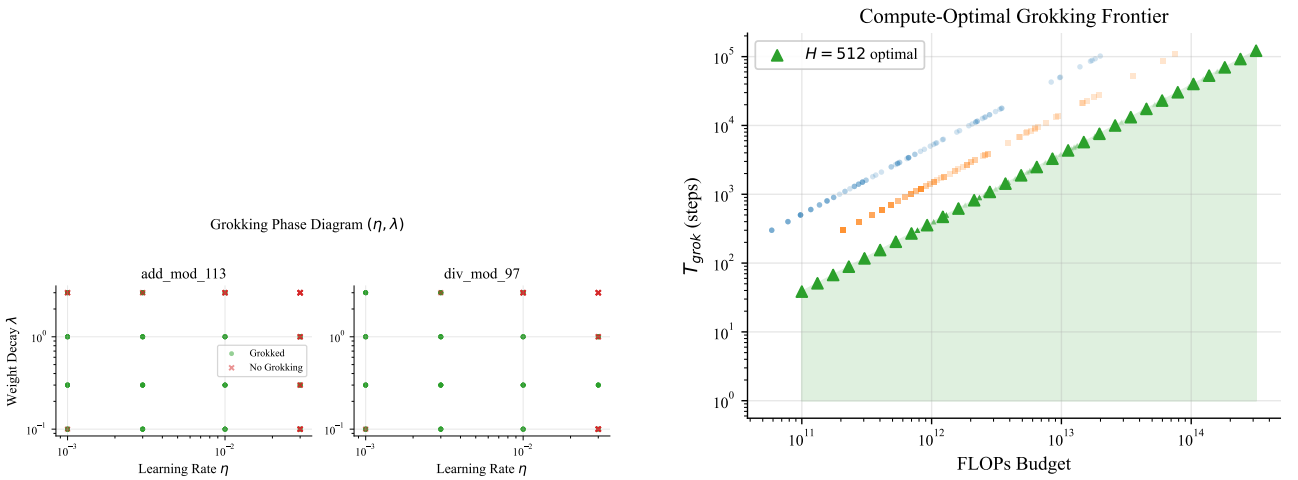

Figure 2. **Left:** Phase diagram in (η, λ) space. Color indicates grokking rate across dataset fractions and widths. At $\lambda \geq 1.0$, nearly all configurations grok; at $\lambda = 0.1$, fewer than 60% do. The boundary is sharp, not gradual. **Right:** Compute-optimal frontier. Total FLOPs $= T_{\text{grok}} \times C_{\text{step}}(H)$, where $C_{\text{step}} \propto H^2$. Wider models grok faster in steps but at higher FLOP cost, mirroring Chinchilla-style trade-offs (Hoffmann et al., 2022).

model connecting lazy-to-rich transitions with exponent values. We extend the scaling law framework to predict not final loss but the *timing* of a qualitative regime transition.

Memorization in generative models. Somepalli et al. (Somepalli et al., 2023) and Carlini et al. (Carlini et al., 2023) documented memorization in diffusion models, finding that dataset scale modulates replication rates. Our $D^{-2.04}$ exponent quantifies an analogous data-complexity effect on memorization persistence.

7. Discussion

The data complexity result. The $D^{-2.04}$ exponent, nearly eight times the capacity exponent $H^{-0.27}$, suggests that the memorization-to-generalization transition is fundamentally governed by how well the training data constrains the solution space, not by how many parameters are available. To halve T_{grok}, one can either double width (reducing T_{grok} by $1.2\times$) or increase data fraction by 41% (reducing T_{grok} by $2\times$). Data is the cheaper lever.

Connections to generative models. Although our experiments use MLPs on modular arithmetic, the memorization-to-generalization transition we characterize is relevant to generative model training. Diffusion models memorize training images at early stages before generalizing to novel samples (Somepalli et al., 2023; Carlini et al., 2023), and dataset scale modulates replication rates. Our scaling methodology — hyperparameter sweep plus power-law fit to transition timing — offers a template for characterizing these transitions in larger-scale generative models. The key prediction is structural: data complexity should dominate capacity in determining transition timing, regardless of architecture. Testing this prediction on transformers, which also grok on modular arithmetic (Nanda et al., 2023; Power et al., 2022), is a natural next step; quantitative data-fraction exponents for attention architectures remain to be measured.

Limitations. (1) The scaling law is fitted on two-hidden-layer MLPs on modular arithmetic; exponents may differ for transformers or natural-language tasks. (2) The width exponent (-0.27 ± 0.10) is estimated from only three discrete levels. (3) $R^2 = 0.732$ (base model) leaves 27% variance unexplained, likely from initialization noise ($\sim$2%), task-specific structure ($\sim$3%), and unswept hyperparameters ($\sim$13%). (4) The phase boundary (Conjecture 1) is empirically observed, not derived from first principles.

Open questions. (i) Does the D^{-2} exponent generalize to other group-structured tasks, or is it specific to modular arithmetic? (ii) Can the phase boundary at λ^* be derived from a PAC-Bayes or minimum description length argument about the relative complexity of the memorizing vs. generalizing solutions? (iii) Is the norm compression trajectory (Conjecture 2) a Lyapunov function for the memorizing regime, with λ controlling the rate of descent? These questions connect the empirical findings reported here to the theoretical foundations that the FoGen community is well-positioned to develop.

Code availability. Code and run logs for all 384 configurations will be released upon publication.

References

B. Barak, B. L. Edelman, S. Goel, S. Kakade, E. Malach, and C. Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *NeurIPS*, 2022.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

B. Bordelon, A. Atanasov, and C. Pehlevan. A dynamical model of neural scaling laws. In *ICML*, 2024.

N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, 2023.

L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. In *NeurIPS*, 2019.

J. Hoffmann, S. Borgeaud, A. Mensch, et al. Training compute-optimal large language models. In *NeurIPS*, 2022.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

Z. Liu, E. J. Michaud, and M. Tegmark. Omnigrok: Grokking beyond algorithmic data. In *ICLR*, 2023.

K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. In *ICLR*, 2020.

W. Merrill, N. Tsilivis, and A. Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. *arXiv:2303.11873*, 2023.

P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt. In *ICLR*, 2020.

N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. In *ICLR*, 2023.

A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *ICLR Workshop on Mathematics of Deep Learning*, 2022.

G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *CVPR*, 2023.

B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. Kernel and rich regimes in overparametrized models. In *COLT*, 2020.

A. Robustness and Variance Analysis

Seed variance. We rerun 10 configurations spanning the T_{grok} range with 5 random seeds each. The median within-configuration CV is 8%. The maximum CV is 18% for a configuration near the grokking phase boundary ($\lambda = 0.1$, $\eta = 0.001$), where stochastic fluctuations can determine whether the network grokks within the 150K step budget. Initialization noise accounts for roughly 1–2% of variance in $\log T_{\text{grok}}$.

Threshold sensitivity. The D^{-2} exponent is stable across generalization thresholds: -2.15 at 85%, -2.05 at 90%, -2.04 at 95%, -1.88 at 99%. The width exponent varies more (-0.28 to -0.19), reflecting its weaker signal.

Cross-validation. Leave-one-level-out cross-validation yields $R^2_{\text{CV}} = 0.67$–0.76 across held-out hyperparameter levels, with median multiplicative prediction errors of 1.4–1.6×. The hardest level to extrapolate is D ($R^2_{\text{CV}} = 0.67$) because the D^{-2} power law must extrapolate over a wider dynamic range.

B. Interaction Model Details

The full interaction model is:

$$\begin{aligned}\log T_{\text{grok}} = &-0.27 \log H - 2.04 \log D - 0.50 \log \eta \\ &- 0.64 \log \lambda + 0.50 \log D \cdot \log \eta \\ &+ 0.35 \log H \cdot \log \eta + 0.38 \log D \cdot \log \lambda \\ &- 0.23 \log H \cdot \log \lambda + (\text{minor terms}) + \varepsilon.\end{aligned}$$

The two remaining interactions ($\log H \times \log D$ and $\log \eta \times \log \lambda$) are not significant ($|t| < 2$).

C. Compute-Optimal Frontier

The per-step FLOP cost for a two-hidden-layer MLP with width H, input dimension $2H$ (embeddings), and C output classes is: $C_{\text{step}}(H) = 2(2H^2 + H^2 + CH) = 2H(3H + C)$. For our architectures ($C \in \{97, 113\}$, $H \in \{128, 256, 512\}$), the H^2 terms dominate: $C_{\text{step}} \approx 6H^2$. Total FLOPs to grok: $F = T_{\text{grok}} \times C_{\text{step}}(H) \propto H^{-0.27} \times H^2 = H^{1.73}$. Since total FLOPs grow with width ($1.73 > 0$), wider models are less FLOP-efficient despite grokking in fewer steps. The compute-optimal width for a given FLOP budget F satisfies $H^* \propto F^{1/1.73} \approx F^{0.58}$, meaning roughly 58% of additional budget should go to width.

D. Falsifiable Predictions

The scaling law and norm dynamics generate three testable predictions:

1. **Fourier mode onset scales as D^{-2}.** If the D^{-2} exponent reflects data coverage requirements for circuit formation, then the step at which specific Fourier features emerge in the weight matrices (measurable via the methodology of Nanda et al. (Nanda et al., 2023)) should scale as D^{-2} when data fraction is varied.

2. **Wider models form higher-rank memorization solutions.** If the weak width exponent ($H^{-0.27}$) reflects wider models building more complex memorization solutions, then the singular value spectrum of weight matrices at T_{mem} should have higher effective rank for larger H.

3. **Norm compression rate predicts T_{grok} better than absolute norm.** The derivative $\|d\theta/dt\|$ at T_{mem} should correlate more strongly with T_{grok} than $\|\theta(T_{\text{mem}})\|$ does. Our data shows $\rho = 0.06$ for the absolute norm; we predict $|\rho| > 0.5$ for the compression rate.