

Grokking as Manifold Discovery

Jin Yanyan, Zhao Lei

Contents

Grokking as Manifold Discovery: A Geometric Reinterpretation of Delayed Generalization	1
1. Introduction: Why Grokking Matters	1
2. Review of Existing Theories	2
2.1 Goldilocks Zone Theory (Liu et al. 2022)	2
2.2 Softmax Collapse Theory (Prieto et al. 2025)	2
2.3 Lazy \rightarrow Rich Transition Theory (Kumar et al. 2024)	3
2.4 Weight Efficiency Hypothesis (Varma et al. 2023)	3
2.5 Mechanistic Interpretability Perspective (Nanda et al. 2023)	3
2.6 Common Blind Spot of Existing Theories	4
3. Unified Framework: The Manifold Discovery Hypothesis	4
3.1 Geometric Interpretation of Memorization vs. Generalization	4
3.2 Geometric Role of Weight Decay	5
3.3 Geometric Interpretation of Softmax Collapse	5
3.4 Geometric Interpretation of Lazy \rightarrow Rich	6
3.5 Geometric Interpretation of Weight Efficiency	6
3.6 A Set of Formalizable Mathematical Statements (Turning “Metaphors” into Provable Propositions)	6
3.7 A Formulation Closer to Real LLMs (GPT-4 Class Systems): Continuous Approximation + Spectral/Low-Complexity Bias	8
4. Reinterpreting Existing Findings	9
4.1 Why Weight Decay Being Too Large/Small Both Fail	9
4.2 Why Data Amount Affects Grokking	9
4.3 Why Over-parameterized Models Are More Prone to Grokking	10
4.4 Why Grokking Is Sudden (and Why It Oscillates)	10
5. Testable Predictions	10
5.1 Intrinsic Dimension Discontinuity	10
5.2 Topological Structure of Representations	11
5.3 Attention Entropy Dynamics	11
5.4 Effect of Forced Rank Constraints	11
6. Experimental Validation	12
6.1 Experimental Configuration	12
6.2 Experiment Group 1: Modular Addition Results	13
6.3 Experiment Group 2: Modular Multiplication Results	15
6.4 Coset Structure Verification: 12 Clusters = $k \bmod 12$	16
6.5 Hypothesis Revision: Two-Stage Model	16

6.6 Adjacency Analysis: Differences in Topology Preservation	17
6.7 Multi-Seed Stability Verification	17
6.8 Cross-Operation Comparison of Two Experimental Groups	18
6.9 Experiment Group 3: Nested Grokking—Topological Possession and the Capacity Hypothesis	19
7. Discussion	21
7.1 Methodological Limitations of Existing Research	21
7.2 Value of the Internal Perspective	22
7.3 Theoretical Significance of Experimental Findings	22
7.4 Limitations and Future Directions	23
8. Conclusion	23
References	24

Grokking as Manifold Discovery: A Geometric Reinterpretation of Delayed Generalization

Authors: Jin Yanyan (lmxxf@hotmail.com), Zhao Lei (zhaosanshi@gmail.com)

Abstract: Grokking—the phenomenon where neural networks suddenly generalize after prolonged overfitting—has accumulated multiple theoretical explanations since its discovery in 2022: Goldilocks Zone, Softmax Collapse, Lazy-Rich transition, etc. This paper reviews these theories and identifies their common blind spot: **most focus on external measurements, lacking direct characterization of representation space geometry**. Among them, the Goldilocks Zone theory touches on the “physical laws” of high-dimensional space and carries substantial theoretical value. We propose a unified framework—the **Manifold Discovery Hypothesis**: memorization is a high-dimensional jagged curve passing through all training points, generalization is discovering the low-dimensional manifold on which data is distributed, and Grokking is the transition from the former to the latter (possibly accompanied by critical state oscillations). **We provide evidence supporting this hypothesis on two experimental groups: modular addition and modular multiplication**: we observed significant drops in effective dimensionality of representations ($78 \rightarrow 8/89 \rightarrow 11$ under $PCA95(k \bmod 12)$ cosets, purity 99.4%), which prompted us to revise the hypothesis into a **two-stage model**: local manifold discovery \rightarrow global gluing. Furthermore, **nested Grokking experiments** reveal the nature of capacity-topology competition: a small model (2-layer, 128-dim) under strong regularization undergoes **topological possession**—the outer \mathbb{Z}_{12} structure is replaced by inner stride=4 structure (test_acc=100% throughout), while scaling up to 4-layer, 256-dim causes the model to fail to Grok entirely (test_acc=51.75%), demonstrating that capacity and regularization pressure must be matched. In one sentence: **high-dimensional curve \rightarrow low-dimensional surface**.

1. Introduction: Why Grokking Matters

In 2022, Power et al. at OpenAI discovered a counterintuitive phenomenon: small Transformers trained on modular arithmetic tasks would first **perfectly overfit** the training set (training loss drops to zero, test accuracy near random), then **tens of thousands or even hundreds of thousands of steps later**, test accuracy suddenly jumps from random level to nearly 100%.

They named this phenomenon **Grokking** (sudden understanding).

This discovery is important because it challenges a core assumption of deep learning:

1. **Classical assumption:** Overfitting is the enemy of generalization; once overfitting occurs, early stopping should be applied
2. **Grokking counterexample:** Overfitting can persist for a long time, then suddenly generalize

If generalization can truly occur after overfitting, the “early stopping” strategy may have killed many models that could have generalized.

The deeper question: **What happens inside the model during Grokking?**

Over the past three years, academia has accumulated multiple theoretical explanations. This paper’s task is to review these theories, identify their blind spots, and propose a unified framework.

2. Review of Existing Theories

2.1 Goldilocks Zone Theory (Liu et al. 2022)

Core idea: Weight norm must fall within a “just right” interval for generalization.

Liu et al. at NeurIPS 2022 found that weight space contains a **hollow spherical shell**, which they called the Goldilocks Zone:

- Radius too large ($\|w\| > w_c$): Overfitting, memorizing training set
- Radius too small ($\|w\| < w_c$): Underfitting, learning nothing
- Just right on the shell ($\|w\| \approx w_c$): Generalization

Grokking mechanism: 1. Large initialization places the model outside the shell 2. Model quickly overfits first (training loss drops to zero) 3. Weight decay slowly pulls weight norm back to Goldilocks Zone 4. Once inside the shell \rightarrow sudden generalization \rightarrow Grokking

The true value of this paper: It suggests that high-dimensional space has its own “physical laws”—weight decay is gravity, Goldilocks Zone is the stable orbit. This is the foundation for all subsequent theories.

Limitation: Describes “where generalization happens,” but doesn’t explain “why it can generalize there.” What is Goldilocks Zone a proxy for?

2.2 Softmax Collapse Theory (Prieto et al. 2025)

Core idea: Without weight decay, Grokking is killed by floating-point precision.

To minimize cross-entropy loss, the model aggressively amplifies the correct answer’s logit (e.g., correct class = 1000, others = 1). When computing softmax, e^{1000} directly overflows, gradients become zero, training stalls.

Role of weight decay: Continuously pulls weights back, preventing infinite logit growth, keeping gradients alive.

Alternative solution: The paper proposes StableMax + orthogonal gradients (preventing gradients from going in the “amplify logit” direction), which can trigger Grokking without weight decay. However, this method may have lower convergence efficiency—weight decay is global compression

with broad scope; orthogonal gradients are local constraints with narrow scope. In practice, weight decay remains the more common choice.

Contribution: Explains “what happens without weight decay.”

Limitation: Only explains “why training doesn’t stop,” not “why it eventually generalizes.”

2.3 Lazy → Rich Transition Theory (Kumar et al. 2024)

Core idea: Grokking is a phase transition from lazy training to feature learning.

Borrowing Neural Tangent Kernel (NTK) language:

- **Lazy regime:** Weights barely move, model acts like linear classifier
- **Rich regime:** Weights adjust significantly, learning true nonlinear features

Grokking occurs at the **phase transition point** from lazy → rich.

Controversy: This camp claims that under specific conditions (shallow networks + MSE loss), Grokking can be triggered without weight decay.

Assessment: The main contribution of this theory is introducing the lazy/rich conceptual framework, though there remains room for explaining “why the phase transition occurs.”

2.4 Weight Efficiency Hypothesis (Varma et al. 2023)

Core idea: Weight decay favors solutions with smaller weights, and generalizing solutions are typically more weight-efficient than memorizing solutions.

- Memorization solution: Requires large weights to hard-memorize each sample
- Generalization solution: Uses concise rules to cover all samples, smaller weights
- Weight decay → penalizes large weights → favors generalization solution

Assessment: Similar to 2.3, this theory provides a useful perspective (small weights ↔ generalization), but is essentially another description of the same phenomenon, not yet revealing causal mechanisms.

2.5 Mechanistic Interpretability Perspective (Nanda et al. 2023)

Nanda et al. at ICLR 2023 (Oral) conducted solid work: **completely reverse-engineered** the algorithm the model learned.

Core finding: Modular arithmetic $(a + b) \bmod p$ is essentially a **cyclic group**—0, 1, 2, ..., $p-1$ connected end to end, forming a discrete ring. The model decomposes this modular operation into Fourier series. Human post-hoc analysis of weight matrices found it equivalent to Fourier transform structure. The model itself just does matrix multiplication, knowing nothing about Fourier formulas.

Discussion: Modular arithmetic is naturally periodic, and expanding with Fourier series is a mathematical tool from 200 years ago. From this perspective, the model “discovering” Fourier structure is more like an inevitable response to the task’s inherent periodicity, rather than an unexpected finding.

Contribution: Hard work, opened up the model to look inside.

Limitation: Doesn't explain why weight decay + overfitting + continued training = Fourier transform.

2.6 Common Blind Spot of Existing Theories

Theory	Question Asked	Question Not Asked
Goldilocks Zone	What weight norm interval	What's special about that interval
Softmax Collapse	Why training doesn't stop	Why it eventually generalizes
Lazy → Rich	How weights change	How representations change
Weight Efficiency	Which solution has smaller weights	Why small weights = generalization
Mechanistic Interp.	What circuit was learned	Why this circuit

Assessment: The Goldilocks Zone theory touches on the “physical laws” of high-dimensional space and carries substantial theoretical value. Softmax Collapse, Lazy→Rich, and Weight Efficiency each provide useful perspectives, but primarily remain at the level of external measurements (weight norm, gradient magnitude, loss curves), not yet revealing the geometric essence of representation space. Mechanistic Interpretability starts looking inside, but focuses on specific circuits, not geometric structure.

Overall, existing theories have done substantial work in “describing the phenomenon,” but there remains room for improvement in “explaining the mechanism.” This paper attempts to provide a complementary unified framework from the geometric perspective.

3. Unified Framework: The Manifold Discovery Hypothesis

We propose a unified framework: **Grokking is a transition from high-dimensional jagged curves to low-dimensional smooth manifolds (possibly accompanied by critical state oscillations).**

3.1 Geometric Interpretation of Memorization vs. Generalization

Memorization = Jagged Curve

When a model overfits the training set, it uses a complex jagged curve to pass through every training sample point. This curve can precisely hit all training data, but it **has no pattern**—just forcibly stringing all points together, with no structural relationship between points.

Generalization = Manifold Discovery

When the model truly “understands” the task, it discovers that training samples are actually distributed on a **low-dimensional manifold** (low-dimensional relative to the model's hidden dim).

Taking modular arithmetic $a+b \bmod p$ as an example: - Input space is p^2 discrete points - But output only depends on $(a+b) \bmod p$, i.e., the **congruence class** - The true structure is a one-dimensional **cyclic group** \mathbb{Z}_p (only one degree of freedom: position)

Generalization means: the model discovered this cyclic group structure, rather than hard-memorizing p^2 input-output pairs.

Grokking = Phase Transition from Curve to Manifold

What happens during Grokking: 1. Before: A high-dimensional jagged curve passing through all points in representation space (memorization solution) 2. After: The curve collapses onto a low-dimensional manifold, whose topological structure corresponds to the task’s true structure (generalization solution)

This is a **phase transition in a multi-stable system**—memorization and generalization solutions are two stable states, and the model jumps from the former to the latter during training. However, note: what we observed experimentally is not a single clean phase transition, but **critical state competition**—the model may repeatedly jump between two solutions (see Section 6.2.3 on oscillation phenomena) before finally stabilizing on the generalization solution.

In one sentence: high-dimensional curve \rightarrow low-dimensional surface.

3.2 Geometric Role of Weight Decay

In this framework, weight decay’s role becomes clear:

Weight Decay = Centripetal Force that Destabilizes Jagged Curves

Without weight decay: - Gradient descent pushes the model toward “lowest loss” - For over-parameterized models, this is “perfectly memorize each training sample” - The model stabilizes on solutions formed by jagged curves

With weight decay: - There’s a force opposite to loss gradient, continuously pulling weights toward “smaller” - Jagged curves need “large weights” to maintain (each inflection point needs dedicated neurons) - Smooth manifolds need “small weights” (structure sharing) - Weight decay **favors manifold encoding**

The True Meaning of Goldilocks Zone

Goldilocks Zone is not “some weight norm interval,” but **activation patterns that can perceive manifold structure.**

- Weights too large: Jagged curve stable, can’t see manifold
- Weights too small: Signal too weak, can’t see any structure
- Just right: Jagged curve unstable + signal strong enough \rightarrow manifold emerges

3.3 Geometric Interpretation of Softmax Collapse

The direct trigger for Softmax Collapse is numerical stability (logits too large cause overflow/underflow/pathological gradients). In this paper’s framework, it can also be understood as a “representation/decision over-unidirectionalization” collapse: when a certain direction is infinitely amplified, distinguishable signals in other directions are drowned out, and exploration is terminated early.

When a direction is over-strengthened: - That direction’s logit tends toward infinity - Other directions’ “voices” are drowned out - Model loses ability to explore other possibilities

This is why Softmax Collapse kills Grokking: Discovering manifolds requires simultaneously “seeing” multiple directions, while collapse shrinks the field of view to a single point.

Role of weight decay: Maintain signal balance across multiple directions, allowing the model to continue exploring.

3.4 Geometric Interpretation of Lazy \rightarrow Rich

Lazy \rightarrow Rich transition is not about weight dynamics, but about **representation complexity**.

- Lazy regime: Representation is linear function of input, can only encode linearly separable structures
- Rich regime: Representation is nonlinear function of input, can encode arbitrary manifolds

Grokking requires Rich regime: Because the true task structure (like cyclic groups) is nonlinear.

Why does Grokking often occur late in training? Because entering Rich regime requires sufficient weight change, and at initialization the model is in Lazy regime.

3.5 Geometric Interpretation of Weight Efficiency

The causal chain of “small weights = generalization”:

1. Smooth manifolds are more “compact” than jagged curves (lower dimension)
2. Compact encoding requires fewer weights to implement
3. Weight Decay favors small weights \rightarrow favors compact encoding \rightarrow favors manifolds

Weight efficiency is the **result** of manifold discovery, not the **cause**.

3.6 A Set of Formalizable Mathematical Statements (Turning “Metaphors” into Provable Propositions)

This section does not attempt to “prove Grokking must happen” (that would require very strong assumptions about networks, data distributions, and optimization dynamics). Instead, we rewrite the key intuitions from this paper’s framework into **mathematical propositions that are provable/verifiable under standard assumptions**:

Note on AdamW and L2 Regularization: The following derivations are based on standard gradient descent + L2 regularization. The AdamW used in experiments is **decoupled weight decay**, whose typical update form can be written as

$$w_{t+1} = (1 - \eta\lambda)w_t - \eta \cdot \text{Adam}(\nabla\mathcal{L}(w_t)),$$

which is not exactly equivalent to “optimizing $\mathcal{L} + \frac{\lambda}{2}\|w\|^2$ ” under adaptive learning rates. Therefore, the mathematical statements in this section should be understood as **theoretical bridges**—they characterize the qualitative effects of weight decay (centripetal force, small weight bias), but cannot be directly applied to AdamW’s precise dynamics.

- 1) **The dynamical form of weight decay** (it really is a centripetal force);
- 2) **The minimum norm bias of L_2 regularization** (it really favors “structure-sharing/low-complexity” interpolating solutions);
- 3) **Why intrinsic dimension drops when “manifold/group structure is discovered”** (at least when “representation depends only on group invariants,” this can be rigorously derived).

Notation: Parameters $w \in \mathbb{R}^m$, empirical loss $\mathcal{L}(w) = \frac{1}{n} \sum_{i=1}^n \ell(f_w(x_i), y_i)$, weight decay coefficient $\lambda \geq 0$, learning rate $\eta > 0$.

Lemma 3.1 (Weight decay = centripetal force in discrete dynamics) Consider the objective with L_2 regularization:

$$J(w) = \mathcal{L}(w) + \frac{\lambda}{2} \|w\|_2^2.$$

Gradient descent on J :

$$w_{t+1} = w_t - \eta \nabla J(w_t) = (1 - \eta\lambda)w_t - \eta \nabla \mathcal{L}(w_t).$$

Thus, the optimization update consists of two superimposed parts: one shrinks w_t proportionally (contraction toward origin), the other descends along the loss gradient. **Proof:** Directly expand $\nabla(\frac{\lambda}{2} \|w\|^2) = \lambda w$. \square

This lemma turns Section 3.2’s “centripetal force” from metaphor into an explicit dynamical term: present at all steps, independent of specific data.

Proposition 3.2 (Linear regression: regularized interpolating solution tends to minimum norm solution) Let $f_w(x) = w^\top x$, squared loss $\ell = \frac{1}{2}(f_w(x) - y)^2$, data matrix $X \in \mathbb{R}^{n \times d}$ with full row rank, and $y \in \mathbb{R}^n$. The optimal solution with L_2 regularization (ridge regression) satisfies

$$w_\lambda = \arg \min_w \frac{1}{2n} \|Xw - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2 = (X^\top X + n\lambda I)^{-1} X^\top y.$$

If an interpolating solution exists (i.e., $\exists w : Xw = y$), then as $\lambda \downarrow 0$, w_λ converges to the **minimum L_2 norm interpolating solution**

$$w_* = \arg \min_{Xw=y} \|w\|_2,$$

and w_* can be written as $w_* = X^\top (XX^\top)^{-1} y$. **Proof (key points):** Use first-order optimality conditions to get closed-form solution. For $\lambda \rightarrow 0$, use X having full row rank to guarantee XX^\top is invertible, and derive using pseudo-inverse limits or equivalent Lagrange multipliers. \square

This proposition clarifies one thing in the simplest convex case: **L_2 regularization doesn’t just “prevent overfitting”; when interpolation is feasible, it selects the “minimum norm” solution.** Replacing “norm” with more general function space norms (like RKHS/Sobolev) naturally yields “smoother/lower frequency” bias—consistent with this paper’s “curve (high frequency) \rightarrow manifold (low frequency)” intuition.

Proposition 3.3 (Separable classification: without regularization weight norm tends to infinity; with regularization optimal solution is bounded) Let binary classification linear model $f_w(x) = w^\top x$, logistic loss $\ell(w; x, y) = \log(1 + \exp(-y w^\top x))$, and data be linearly separable: $\exists \bar{w}$ such that for all i , $y_i \bar{w}^\top x_i > 0$. Then: 1) Without L_2 regularization ($\lambda = 0$), the infimum of minimum empirical loss is 0, but generally **no finite-norm optimal solution exists** (can scale w along the separable direction to make loss arbitrarily close to 0). 2) With L_2 regularization ($\lambda > 0$), the objective $J(w) = \frac{1}{n} \sum_i \ell(w; x_i, y_i) + \frac{\lambda}{2} \|w\|^2$ is **strongly convex + lower semi-continuous** on \mathbb{R}^d , thus there exists a unique optimal solution w_λ , with $\|w_\lambda\| < \infty$. **Proof (key points):** 1) Use $\ell(\alpha w) \rightarrow 0$ ($\alpha \rightarrow \infty$) to show loss can be pushed arbitrarily small but not achieved. 2) The L_2 term provides strong convexity and coerciveness: $\|w\| \rightarrow \infty$ implies $J(w) \rightarrow \infty$, thus a unique minimum exists. \square

This proposition corresponds to Section 2.2’s Softmax/Logit explosion phenomenon: in cross-entropy/logistic loss, the shortcut to “keep improving” is often to amplify the margin to infinity; weight decay turns it into a bounded optimization problem, and gradients won’t naturally vanish due to “infinite norm” escape.

Proposition 3.4 (“Discovering group structure” \Rightarrow representation degrees of freedom decrease: a rigorously testable sufficient condition) Using modular addition as an example, inputs are $(a, b) \in \mathbb{Z}_p^2$, let sum be $s = a + b \pmod{p} \in \mathbb{Z}_p$. Let some layer’s representation be $h(a, b) \in \mathbb{R}^k$, and there exists a function $\phi : \mathbb{Z}_p \rightarrow \mathbb{R}^k$ such that

$$h(a, b) = \phi(s) \quad \text{depends only on } s = (a + b) \pmod{p}.$$

Then the representation set $\{h(a, b) : a, b \in \mathbb{Z}_p\}$ contains at most p points (exactly $\phi(\mathbb{Z}_p)$), and its effective degrees of freedom are no longer proportional to p^2 , but bounded by p . Furthermore, if there exists a smooth embedding $\Phi : S^1 \rightarrow \mathbb{R}^k$ and a homomorphism $\iota : \mathbb{Z}_p \hookrightarrow S^1$ (embedding the discrete cyclic group into the circle) such that $\phi = \Phi \circ \iota$, then these representation points lie on a one-dimensional manifold (circle). **Proof:** The first part follows immediately from “function depends only on s ”: different (a, b) with the same congruence class map to the same h , so the number of distinct representations is bounded by the number of congruence classes p . The second part is direct substitution into the composition mapping definition: $\phi(\mathbb{Z}_p) \subseteq \Phi(S^1)$. \square

This proposition provides an **operationally testable sufficient condition** for Section 5.1’s “intrinsic dimension discontinuity”: once some layer’s representation truly “forgets (a, b) ’s two degrees of freedom, keeping only s ’s one degree of freedom,” then the degrees of freedom decrease you see using PCA/TwoNN/local dimension estimation is not metaphysics, but forced by representation factorization.

3.7 A Formulation Closer to Real LLMs (GPT-4 Class Systems): Continuous Approximation + Spectral/Low-Complexity Bias

If betting on “which mathematical path real-world GPT-4 more resembles,” I would choose: **continuous approximation (S^1 / low-dimensional manifold) + spectral/low-complexity bias**, rather than doing completely rigorous algebraic derivations only in discrete finite groups. The reason is simple: real LLM training data and task distributions are closer to “samples from a continuous world,” and Transformer representations typically exhibit strong “learn low frequency/simple structure first, then high frequency/exceptions” dynamical bias.

Below is a version you can directly write into papers, requiring readers to know less abstract algebra (it’s compatible with Section 3.6, just replacing “discrete group \mathbb{Z}_p ” with “continuous circle S^1 approximation”).

Setup (viewing \mathbb{Z}_p as evenly-spaced samples on S^1) Map $k \in \mathbb{Z}_p$ to angle $\theta_k = 2\pi k/p \in [0, 2\pi)$, and map “sum s ” to circle position θ_s . Assume some layer’s representation satisfies the approximate form

$$h(a, b) \approx \Phi(\theta_{(a+b) \pmod{p}}) \quad (\Phi : S^1 \rightarrow \mathbb{R}^d \text{ continuous/smooth}).$$

Then “discovering structure” is equivalent to: the network learns at some layer a **low-dimensional parameterization** $\theta \mapsto \Phi(\theta)$, rather than memorizing an independent representation for each (a, b) .

Proposition 3.5 (Fourier expansion on the circle: low frequency = smooth structure, high frequency = jagged memorization) For each output dimension, let $\Phi_j(\theta)$ be square-

integrable, then there exists a Fourier series

$$\Phi_j(\theta) = \sum_{m \in \mathbb{Z}} c_{j,m} e^{im\theta}.$$

If coefficients decay rapidly in frequency (e.g., $\sum_m m^2 |c_{j,m}|^2 < \infty$), then Φ is smoother in θ ; conversely, if many high-frequency components are needed to fit the fine-grained differences of training points, this corresponds to “jagged/memorization” style representation. **Note:** This is not “proving networks must learn low frequency,” but providing a quantifiable characterization: you can do discrete Fourier transform (DFT) on representations sampled at θ in experiments, checking whether energy spectrum concentrates from high to low frequency before and after Grokking.

Proposition 3.6 (Minimizing smoothness regularization suppresses high frequency: a clean variational conclusion) Consider the variational problem of fitting target function $g(\theta)$ on S^1 :

$$\min_{\Phi} \int_{S^1} \|\Phi(\theta) - g(\theta)\|^2 d\theta + \alpha \int_{S^1} \|\partial_{\theta} \Phi(\theta)\|^2 d\theta, \quad \alpha > 0.$$

The optimal solution satisfies shrinkage for frequency m in Fourier domain: high-frequency components are penalized more strongly (because $\|\partial_{\theta} e^{im\theta}\|^2 \propto m^2$), so the solution is biased toward low-frequency/smooth structure. **Note:** In real networks you don’t have explicit $\int \|\partial_{\theta} \Phi\|^2$ regularization, but “parameter norm/weight decay + optimization dynamics” often exhibits similar low-complexity bias; this gives you a bridge closer to engineering intuition: **weight decay** \rightarrow **low-complexity bias** \rightarrow **high frequency suppressed** \rightarrow **representation smoother** \rightarrow **more like low-dimensional manifold**.

Including this section allows you to explain Section 5’s predictions using “spectral energy shifting from high to low frequency,” and it aligns better with real LLM observations (learn commonalities first, then exceptions).

4. Reinterpreting Existing Findings

4.1 Why Weight Decay Being Too Large/Small Both Fail

Too small: Centripetal force insufficient, model stabilizes on jagged curve, never explores manifold.

Too large: Centripetal force too strong, signal suppressed, can’t even draw jagged curve, let alone discover manifold.

Just right: Jagged curve unstable but signal strong enough, model forced to explore, eventually discovers manifold.

4.2 Why Data Amount Affects Grokking

Too little data: Samples on manifold too sparse, can’t recover manifold structure. Model can only draw jagged curve.

Too much data: Manifold signal too strong, model directly discovers manifold, no overfitting phase, no “delayed generalization.”

Just right: Manifold signal exists but not obvious, model draws jagged curve first, slowly discovers the structure behind the curve.

This explains why Grokking needs a “Goldilocks” data amount.

4.3 Why Over-parameterized Models Are More Prone to Grokking

Over-parameterization = representation space large enough

- Small model: Representation space might not fit the true manifold at all
- Large model: Representation space large enough, manifold can exist, just needs time to discover

The over-parameterization paradox: - Traditional view: Over-parameterization leads to overfitting - Grokking view: Over-parameterization is a **prerequisite** for generalization, because it provides sufficient representation space

Weight decay solves over-parameterization’s “too many degrees of freedom” problem: although the space is large, centripetal force pushes the model toward low-dimensional manifolds.

4.4 Why Grokking Is Sudden (and Why It Oscillates)

Phase transitions are not continuous

Manifold discovery is a **topological event**:

- Before: High-dimensional jagged curve (memorization solution)
- After: Low-dimensional smooth manifold (generalization solution)

The collapse from high to low dimension has no stable intermediate state, so Grokking is sudden.

Analogy: Supercooled Water

A more accurate analogy than the Curie point (unidirectional phase transition) is **supercooled water**—water can remain liquid below 0°C, but random perturbation can trigger instant freezing, or it may melt again.

Grokking is similar: memorization and generalization solutions are two (pseudo-)stable states, weight decay continuously pushes the model toward the critical point, and optimization noise can trigger the transition. This explains the **critical state oscillations** observed experimentally—the model may repeatedly jump between 100% → 0.8% → 100% until finally stabilizing.

5. Testable Predictions

This section is deliberately written to not “close the loop perfectly.” We don’t promise what some curve **must** look like, but give a set of **falsifiable observational signatures**: after you do experiments, results will push the story in some direction (support / revise / refute), rather than falling into the “can be explained either way” word game.

5.1 Intrinsic Dimension Discontinuity

Observational signature: Before and after Grokking, intrinsic dimension of intermediate layer representations may show “discontinuity/collapse,” or may only show “slow drift.”

Two mutually exclusive readings (giving the experiment a “choose one” verdict): - **If representation truly factorizes** (e.g., some layer starts depending approximately only on congruence class/group invariant), intrinsic dimension should drop significantly from “close to sample degrees of freedom” to “close to task degrees of freedom” (like modular addition’s 1D structure). - **If generalization comes**

from other mechanisms (e.g., just decision boundary becoming more stable, but representation not factorizing), you might not see obvious dimension reduction, only smooth variation in dimension estimates during training.

Experimental design: 1. Train model on modular arithmetic, record intermediate layer activations 2. Use existing algorithms (like SVD, PCA, or TwoNN) to estimate intrinsic dimension 3. Plot intrinsic dimension vs. training steps

Decision point: Whether intrinsic dimension curve shows structural inflection (discontinuity or clear turning point) near test accuracy jump.

5.2 Topological Structure of Representations

Observational signature: After Grokking, intermediate layer representations’ topological structure may start “matching” task structure, or may only show weak correlation.

- Modular arithmetic $a + b \bmod p$: Representation should form a one-dimensional ring
- Symmetric group tasks: Representation should form corresponding group manifold

Experimental design: 1. Extract intermediate layer representations before and after Grokking 2. Use persistent homology to compute topological invariants 3. Compare with task’s true topological matching degree

Decision point: Whether Betti numbers/persistence diagrams show clear topological signatures near “grokking” (e.g., ring structure’s β_1 enhancement), and stability across different random seeds.

5.3 Attention Entropy Dynamics

Observational signature: During Grokking, attention pattern entropy may show “low→high→medium” exploration-convergence process, or may completely not follow this narrative (then attention isn’t the key variable).

Two distinguishable patterns: - **Exploration-convergence type:** Early low entropy (specialized patterns) → pre-critical entropy rise (pattern diversification) → afterward falls and stabilizes (shared structure). - **Monotonic/irrelevant type:** Entropy changes monotonically or is very noisy, unrelated to test jump timing—this would weaken “attention collapse”’s explanatory weight for this task.

Experimental design: 1. Record attention matrix at each training step 2. Compute attention distribution entropy 3. Plot entropy vs. training steps

Decision point: Whether entropy peaks/valleys align with test jump, and reproducible across multiple seeds.

5.4 Effect of Forced Rank Constraints

Observational signature: If “representation dimension/rank” is truly the bottleneck, forcing low-rank constraints on intermediate layers should systematically change Grokking timing; if almost no change, then the determining factor may be elsewhere (optimization/numerical stability/normalization, etc.).

- Rank constraint = task’s true degrees of freedom: Accelerate Grokking (directly telling model answer’s dimension)

- Rank constraint $<$ task’s true degrees of freedom: Prevent Grokking (representation space can’t fit manifold)
- Rank constraint $>$ task’s true degrees of freedom: No effect or slight slowdown

Experimental design: 1. Add low-rank bottleneck to intermediate layer (like linear layer limiting output dimension) 2. Vary bottleneck dimension, record Grokking time 3. Compare with no-bottleneck baseline

Decision point: Whether Grokking time vs. bottleneck dimension curve has clear “phase transition region” (once below some threshold, completely fails), and whether threshold is same order of magnitude as task’s minimum degrees of freedom.

6. Experimental Validation

We designed two experimental groups to validate the manifold discovery hypothesis: modular addition ($a + b \bmod 97$) and modular multiplication ($a \times b \bmod 97$). Each group includes five sub-experiments: intrinsic dimension analysis, topological structure analysis, activation dynamics analysis, bottleneck experiment, and manifold visualization.

6.1 Experimental Configuration

Parameter	Modular Addition	Modular Multiplication
Task	$(a + b) \bmod 97$	$(a \times b) \bmod 97$
Dataset size	$97^2 = 9409$ pairs	$(97 - 1)^2 = 9216$ pairs
Training set ratio	30%	30%
Model	2-layer Transformer	2-layer Transformer
Hidden dim	128	128
Attention heads	4	4
Optimizer	AdamW	AdamW
Learning rate	1e-3	1e-3
Weight decay	1.0	1.0
Total steps	150,000	150,000

Evaluation protocol: - Input encoding: (a, b) as two tokens, mapped to hidden_dim via embedding layer - Output: Classification task, 97 classes for addition, 96 classes for multiplication - Loss: Standard cross-entropy - Test set: 70% of data (non-overlapping with training set) - Random accuracy: Addition $1/97 \approx 1.03\%$, multiplication $1/96 \approx 1.04\%$ - Accuracy $< 1\%$ in the text is around random level

Mathematical background: - Modular addition corresponds to additive group \mathbb{Z}_{97} , 97 elements, cyclic group, 97 output classes - Modular multiplication corresponds to multiplicative group \mathbb{Z}_{97}^* , 96 elements (excluding 0), isomorphic to \mathbb{Z}_{96} (also cyclic group), 96 output classes - Both have identical group-theoretic structure (cyclic groups), but modular multiplication is more nonlinear in input coordinates (a, b) —addition only needs to learn $(a + b) \bmod p$, multiplication needs to learn discrete log addition $(\log a + \log b) \bmod (p - 1)$

Analysis method parameters (reproducible configuration):

Analysis Type	Tool/Library	Key Parameters
Persistent homology	ripser + persim	Point cloud subsample 500 points, Euclidean distance, Vietoris-Rips complex, maxdim=1, persistence threshold 0.1, random_state=42
UMAP visualization	umap-learn	n_neighbors=15, min_dist=0.1, metric='cosine', random_state=42
Adjacency analysis	scipy.spatial.distance	k=2 nearest neighbors on cluster centers, Euclidean distance, cluster center = mean of UMAP embeddings for same-label samples
Activation extraction	PyTorch	Last Transformer layer output, average pooling of two tokens, yielding $n \times 128$ matrix

Notes: - Betti numbers from persistent homology are sensitive to subsample size and threshold; values in tables should be understood as orders of magnitude, not precise values - UMAP’s topological preservation depends on parameter choices; different parameters may yield different visual structures - Adjacency analysis is performed in UMAP’s 2D reduced space, not the original 128-dimensional space

6.2 Experiment Group 1: Modular Addition Results

6.2.1 Intrinsic Dimension Discontinuity [OK] **Method description:** - Extract last Transformer layer output activations (all test samples) - Average pooling of two tokens per sample, yielding $n \times 128$ matrix - Use PCA, take minimum number of principal components explaining 95% cumulative variance as “effective dimension” - Note: This is “effective rank/energy dimension,” dependent on normalization method, not equivalent to strict intrinsic dimension estimation (like TwoNN)

Using PCA (95% variance explained threshold) to estimate intrinsic dimension:

Step	PCA Dim	Test Acc	State
1000	78	0.1%	Initial (high-dim chaos)
7000	8	24.9%	Lowest point
9000	12	100%	First Grokking
11000	3	8.6%	Collapse
150000	13	96%	Final stable

Conclusion: Dimension dropped sharply from 78 to 8, **supports hypothesis.**

6.2.2 Topological Structure Analysis [OK] Using persistent homology to compute Betti numbers (step 7000 is the dimension minimum before Grokking, step 14000 is recovery period after first collapse):

Metric	Before (step 7000)	After (step 14000)	Change
β_0 (connected components)	500	6	-99%
β_1 (loops)	504	0	-100%
β_0 max persistence	7.59	0.23	-97%
β_1 max persistence	2.89	0.04	-99%

Conclusion: Under our topological pipeline and threshold, we observed clear “topological collapse” trend (connected components dropped significantly, loop structures largely disappeared), consistent with the narrative of “moving from fragmented memorization solution toward more structured representation.”

***[!]** Important note on $\beta_1=0$: **After Grokking, β_1 dropped from 504 to 0, seemingly contradicting “ring structure.” But this precisely shows the model learned 97 discrete clusters separated by label, not representations distributed continuously along a ring. From persistent homology’s perspective, 97 mutually disconnected clusters do not form a loop ($\beta_1=0$), but they can be understood as “discretely sampled cyclic group”—each cluster corresponds to a congruence class, and adjacency relationships between clusters (in UMAP visualization) still reflect group structure.** This explains why the adjacency score in Section 6.6 is 0%—the model learned classification, not topology.**

6.2.3 Activation Dynamics Analysis [OK] Each time accuracy collapsed, L2 norm and standard deviation dropped simultaneously:

Step	Test Acc	L2 Norm	Std	State
9000	100%	9.9	0.88	Normal
14000	0.8%	1.79	0.16	Collapse
17000	98.6%	9.7	0.87	Recovery
26000	1%	1.88	0.17	Collapse
54000	0.5%	3.65	0.33	Collapse

Unexpected finding: Critical state oscillation—the model repeatedly jumps between generalization solution and “blank state” (L2 drops from ~ 10 to ~ 2 during collapse), not a one-time phase transition.

6.2.4 Bottleneck Experiment [OK] Adding low-rank bottleneck to intermediate layer, testing dimension lower bound:

Bottleneck Dim	Final Acc	State
1	7.6%	[X] Complete failure
2	20.9%	[X] Failure
4	8.8%	[X] Failure
8	1.3%	[X] Failure
16	99.7%	[OK] Slow Grok (2x)
32	100%	[OK] Slower (4x)
64	85%	[?] Unstable

Bottleneck Dim	Final Acc	State
128 (baseline)	100%	[OK] Normal

Conclusion: Critical point is between 8-16 dimensions. **Supports hypothesis**—dimension lower bound exists.

6.2.5 Manifold Visualization [OK] Using UMAP to reduce 128-dim representations to 2D:

- **Step 5,000 (memorization phase):** Chaotic points, mixed colors
- **Step 30,000 (transition phase):** 97 clusters appear, one cluster per label
- **Step 100,000 (post-Grok):** Tighter clusters, stable structure

Conclusion: Transition from chaotic points to label-separated cluster structure is directly observable, supporting the claim that “representations underwent structural reorganization during training.”

6.3 Experiment Group 2: Modular Multiplication Results

6.3.1 Intrinsic Dimension [OK]

Step	PCA Dim	Test Acc
1000	89	0.1%
9000	11	100%
150000	17	100%

Comparison with modular addition: Higher initial dimension (89 vs 78), higher minimum dimension (11 vs 8), consistent with expectation that multiplicative group structure is more complex.

6.3.2 Topological Structure Using persistent homology to compute Betti numbers (same time points as addition: step 7000 vs step 14000):

Metric	Before (step 7000)	After (step 14000)	Change
β_0	500	500	No change
β_1	870	179	-79%
β_1 max persistence	2.30	1.96	-15%

Unexpected finding: Modular multiplication did not show β_0 collapse (unlike addition’s 500→6), but β_1 dropped significantly. This suggests **the two operations have different topological evolution patterns**.

6.3.3 Bottleneck Experiment

Bottleneck Dim	Final Acc	State
1-8	< 28%	[X] Failure
16	98.2%	[?] Close but unstable

Bottleneck Dim	Final Acc	State
32	99.9%	[OK] Slow Grok
64	29.3%	[X] Anomaly!
128	100%	[OK] Normal

Key findings: 1. Higher critical point (16-32 dim vs 8-16 dim) 2. 64-dim anomaly—both experiments unstable (addition 85%, multiplication 29%), possibly some resonance interval

6.3.4 Manifold Visualization

- **Step 5,000:** Two chaotic blobs (different from addition’s single blob)
- **Step 30,000:** 96 points arranged diagonally
- **Step 100,000:** ~12 large clusters (mixed colors)

Key question: Why 12 clusters instead of 96?

6.4 Coset Structure Verification: 12 Clusters = $k \bmod 12$

We verified that the 12 clusters correspond to coset structure of the multiplicative group:

Method: 1. Find primitive root $g = 5$ of 97 2. For each label y , compute discrete log k such that $5^k \equiv y \pmod{97}$ 3. Use KMeans clustering, check $k \bmod 12$ distribution in each cluster

Results:

Cluster	Dominant $k \bmod 12$	Purity
0	11	99.64%
1	4	99.81%
...
10	2	100.00%

Average purity: 99.4%

Conclusion: The model learned quotient group coordinates $k \bmod 12$ (i.e., coset identifiers), not the complete discrete log k . This means: - Through discrete log isomorphism $\mathbb{Z}_{97}^* \cong \mathbb{Z}_{96}$, the multiplicative group is mapped to additive group - On \mathbb{Z}_{96} , $k \bmod 12$ corresponds to quotient group $\mathbb{Z}_{96}/12\mathbb{Z}_{96} \cong \mathbb{Z}_{12}$ - 12 clusters = 12 cosets of 8-order subgroup $H = 12\mathbb{Z}_{96}$ - **Local manifold discovery complete, global gluing incomplete**

6.5 Hypothesis Revision: Two-Stage Model

Based on experimental results, we revise the original hypothesis into a **two-stage model**:

Stage	Phenomenon	Metric Change	Mod Addition	Mod Multiplication
1. Local manifold discovery	Intra-component structuring	β_1 drops, dimension drops	[OK] Complete	[OK] Complete

Stage	Phenomenon	Metric Change	Mod Addition	Mod Multiplication
2. Global gluing	Inter- component alignment	β_0 drops	[OK] Complete	[X] Incomplete

Explanation: - Modular addition: Both stages complete, β_0 dropped from 500 to 6 - Modular multiplication: Only completed first stage, β_0 remains 500, but learned quotient group structure

Intuitive phrasing (optional, narrative style): > Grokking = entropy phase transition + structural crystallization > - Addition = liquefaction (continuous ring) > - Multiplication = crystallization (subgroup crystals)

6.6 Adjacency Analysis: Differences in Topology Preservation

We further verified whether the model learned the group’s topological structure (adjacency relationships), or just discrete equivalence class classification.

Modular addition: Check if 97 clusters’ nearest neighbors are $s \pm 1 \bmod 97$ - Adjacency score: **0%** (random baseline 2.1%) - Conclusion: [X] No evidence of ring adjacency being preserved

Modular multiplication: Check if 12 cosets’ nearest neighbors are $(k \pm 1) \bmod 12$ - Adjacency score: **100%** (random baseline 16.7%) - All 12 cosets’ nearest neighbors are $k \pm 1$ - Conclusion: [OK] Model perfectly learned \mathbb{Z}_{12} ’s ring topology

Comparison:

Metric	Mod Addition	Mod Multiplication
Adjacency score	0%	100%
Random baseline	2.1%	16.7%
Structure type	Discrete equivalence classes	Perfect \mathbb{Z}_{12} ring

Core insight: The multiplicative group’s quotient structure (\mathbb{Z}_{12}) preserved topological integrity in embedding space, while the additive group’s cyclic structure (\mathbb{Z}_{97}) was scattered into discrete points. Possible reason: 97 points are too many, model had insufficient motivation to preserve adjacency; 12 cosets are just in a “tractable” scale.

6.7 Multi-Seed Stability Verification

We used 3 additional seeds (1001, 1002, 1003) to verify reproducibility of core findings.

Modular addition:

Seed	First Grok	Oscillation Count	Final Accuracy
1001	9000	13	100% [OK]
1002	10000	20	29.1% [X]
1003	9000	16	100% [OK]
Mean	9333 ± 471	16.3 ± 2.9	2/3 success

Modular multiplication:

Seed	First Grok	Oscillation Count	Final Accuracy
1001	13000	14	33.3% [X]
1002	10000	17	100% [OK]
1003	10000	12	98.9% [OK]
Mean	11000 ± 1414	14.3 ± 2.1	2/3 success

Core findings:

1. **Grokking success rate ~67%:** Both experiments show 2/3 success, phase transition is not guaranteed
2. **Oscillation is universal:** All seeds show 12-20 oscillations, critical state competition confirmed
3. **Consistent failure mode:** Failed seeds have final accuracy $\approx 30\%$ (training set ratio), indicating stuck in memorization solution

Revised theoretical narrative: > Grokking = high-dimensional curve \rightarrow oscillation \rightarrow low-dimensional manifold > But phase transition is not guaranteed to succeed, with $\sim 1/3$ probability of getting stuck in memorization solution.

6.8 Cross-Operation Comparison of Two Experimental Groups

Metric	Mod Addition	Mod Multiplication	Conclusion
Dimension change	$78 \rightarrow 8$	$89 \rightarrow 11$	Both dropped sharply [OK]
First Grok	9333 ± 471 steps	11000 ± 1414 steps	Multiplication slightly slower
Oscillation count	16.3 ± 2.9	14.3 ± 2.1	Both oscillate
Success rate	67% (2/3)	67% (2/3)	Phase transition is probabilistic
Topological change	β_0 500 \rightarrow 6, β_1 504 \rightarrow 0	β_0 unchanged, β_1 870 \rightarrow 179	Different patterns
Adjacency score	0%	100%	Multiplication preserves topology
Bottleneck critical point	8-16 dim	16-32 dim	Multiplication needs more dimensions
Final structure	97 clusters	12 cosets	Addition=full group, Multiplication=quotient group
64-dim anomaly	85%	29%	Both unstable

Core conclusion: The manifold discovery hypothesis is validated on both operations, but the “depth” and “quality” of discovery differ—modular addition discovered the complete cyclic

group structure but lost topological adjacency, modular multiplication only discovered quotient group structure but perfectly preserved ring topology.

6.9 Experiment Group 3: Nested Grokking—Topological Possession and the Capacity Hypothesis

6.9.1 Motivation Section 6.4 found that the modular multiplication model learned \mathbb{Z}_{12} quotient group structure (12 cosets, adjacency score 100%), but the 8 elements inside each coset (\mathbb{Z}_8 substructure) were memorized by rote. **Can stronger regularization and longer training force the model to also discover \mathbb{Z}_8 ’s internal structure?**

6.9.2 Experimental Configuration

Parameter	Small model (baseline)	Large model (scaled)
Task	$(a \times b) \bmod 97$	Same
Layers	2	4
Hidden dim	128	256
Attention heads	4	8
Parameters	~100K	~800K
Weight decay	1.0 / 1.5 / 2.0 / 5.0	2.0
Total steps	1M (some 5M)	1M

Analysis tool: `scan_strides.py`—checks adjacency scores at stride=1/2/4 separately for the outer layer (12 cosets) and inner layer (8 elements per coset). Stride is “step size”: stride=1 means standard ring ($0 \rightarrow 1 \rightarrow 2 \rightarrow \dots \rightarrow 11 \rightarrow 0$), stride=2 means every-other-hop (evens in one loop, odds in another), stride=4 means every-fourth-hop. Stride1 means the model chose a compressed topological encoding.

6.9.3 Finding 1: Topological Possession Regardless of WD strength (as long as Grokking was triggered), the outer \mathbb{Z}_{12} structure eventually collapsed and was replaced by inner stride=4 structure. **test_acc=100% throughout**—the model’s output remained correct while its internal representation underwent complete reorganization.

Timeline for wd=2.0 (key steps excerpted):

Step	outer_s2	inner_s4	PCA dim	Interpretation
20,000	1.000	0.115	13	Outer stride=2 locked
100,000	1.000	0.208	11	Inner signal emerging
140,000	0.583	0.000	6	Turning point , PCA plummets
200,000	0.083	0.542	13	Inner takes off
860,000	0.000	0.896	12	Inner peak
1,000,000	0.083	0.844	15	Final (still oscillating)

Timeline for wd=1.5: Outer chose stride=1 (standard large ring), survived to ~350K steps before collapse, similarly replaced by inner stride=4.

Conclusion: The small model’s capacity ($\sim 100\text{K}$ parameters) is insufficient to simultaneously maintain two levels of topological structure, forcing a binary choice—first learns the outer layer (coarse classification is easier), later gets possessed by the inner layer (stride=4 encoding is more efficient).

6.9.4 Finding 2: WD Determines Topology Selection

WD	Outer topology choice	Interpretation
1.5	stride=1 (12-step standard large ring)	Moderate WD pressure, model can afford precise structure
2.0	stride=2 (two 6-step small rings)	Heavy WD pressure, chooses lossy compression—splits into two sub-rings, reducing weight coupling

Stride=2 splits \mathbb{Z}_{12} into two \mathbb{Z}_6 sub-rings (evens in one loop, odds in another), representing the model’s **spontaneous symmetry breaking** under WD pressure.

6.9.5 Finding 3: Inner Layer Always Chooses stride=4 = gcd(12,8) Regardless of whether the outer layer chose stride=1 or stride=2, the inner layer always converged to stride=4.

4 = gcd(12,8)—the outer layer cycles every 12 steps, the inner every 8 steps, and every 4 steps both layers’ phases align. This is the mathematically shortest path for encoding both layers of information with the same set of weights. The model autonomously discovered this number-theoretic structure.

6.9.6 Finding 4: WD Non-Monotonicity—Extended Training Is Poison wd=2.0 extended from 1M to 5M steps:

Training steps	train_acc	test_acc	Status
1M	100%	100%	Sweet spot
3.56M	—	3.4%	Collapse
5M	78.2%	73.4%	Permanent degradation

Weight Decay first forces structure to emerge (Grokking), then destroys it (over-compression). An optimal training length exists; beyond it, WD becomes poison.

6.9.7 WD Phase Diagram

WD	test_acc (1M steps)	Behavior
1.0	89.6%	No Grok, mostly memorization
1.5	100%	Grok, outer s1→350K collapse→inner s4 takes over
2.0	100%	Grok, outer s2→140K collapse→inner s4 takes over
5.0	77.7%	Over-compression, can’t even memorize

The stronger the WD, the faster the outer layer dies (350K \rightarrow 140K), but the final outcome is the same—only inner stride=4 survives.

6.9.8 Finding 5: Scaling Experiment—Simple Capacity Hypothesis Does Not Hold If topological possession is due to insufficient capacity, then scaling the model by 8x (4-layer, 256-dim, ~800K parameters) should allow two topological levels to coexist.

Result: train_acc=56.4%, test_acc=51.75%—**couldn’t even fit the training set, let alone Grok.**

Key scan_strides data:

Step	outer_s1	inner_s4	Interpretation
360,000	1.000	0.000	Outer s1 perfect flash (survived only 10K steps)
380,000	0.000	1.000	Inner s4 perfect flash (survived only 10K steps)
400,000+	0.000	0.1-0.4	Complete collapse, aimless drifting
1,000,000	0.000	0.177	Final: no stable topology

Comparison with small model:

	Small model (2L, 128d)	Large model (4L, 256d)
test_acc	100%	51.75%
Outer lock duration	100-350K steps	Only 10K steps (360K flash)
Inner stable score	0.5-0.9	No stability (drifting)
Final topology	Inner stride=4 stable	No stable topology

Conclusion:

Not “bigger house enables coexistence,” but “bigger house can’t find the walls.”

- Small model: Low capacity → WD=2.0 effectively constrains → forces structure → possession
- Large model: High capacity → WD=2.0 pressure is relatively insufficient → parameter space too large, optimization landscape too flat → no topology can stabilize

The large model momentarily touched perfect topology at steps 360K and 380K (outer_s1=1.0 and inner_s4=1.0), indicating the target structure **exists** in the solution space, but the optimizer cannot find a stable path to it.

Conjecture: Capacity and regularization pressure must be matched—the large model likely needs much stronger WD (e.g., 5.0-10.0) to force structure emergence. This is consistent with findings from the Epiplexity learnability experiments: bigger better, unless pressure is simultaneously increased.

7. Discussion

7.1 Methodological Limitations of Existing Research

Measurement targets of existing Grokking research: - Weight norm - Gradient magnitude - Loss curves - Test accuracy

These are all **external observables**—values measurable from outside the model.

Analogy: Studying human learning by only measuring brainwaves and pupil diameter, not asking “what does learning feel like.”

This methodology can answer “under what conditions Grokking occurs,” but cannot answer “what Grokking is.”

7.2 Value of the Internal Perspective

The manifold discovery hypothesis proposed in this paper is an **internal perspective** explanation:

- Not asking “what weight norm interval”
- Asking “what is the structure of representation space”

The value of this perspective: 1. **Unification:** Can simultaneously explain Goldilocks Zone, Softmax Collapse, Lazy→Rich, and is compatible with Nanda et al.’s circuit findings 2. **Predictability:** Generates verifiable experimental predictions (intrinsic dimension, topological structure, etc.) 3. **Heuristic value:** Points to new research directions (directly measuring geometric properties of representation structure, not just reverse-engineering specific circuits)

7.3 Theoretical Significance of Experimental Findings

The experimental validation in this paper brings several important theoretical insights:

1. Critical State Oscillation

The original hypothesis predicted Grokking is a one-time topological phase transition. But experiments observed the model repeatedly jumping between generalization solution and “blank state” (accuracy $100\% \rightarrow 0.8\% \rightarrow 100\% \rightarrow \dots$). This suggests Grokking is more like “supercooled water”—unstable between ice and water, random perturbation can trigger crystallization or melting.

2. 64-Dimension Anomaly

Both experimental groups showed anomalous behavior at 64-dim bottleneck (addition 85%, multiplication 29%). One explanation is that 64-dim is in a “overcomplete but unaligned” critical state—too small forces you to Grok, too large lets you Memorize, middle is schizophrenia.

3. Discovery of Quotient Group Structure

In the modular multiplication experiment, the model only learned $k \bmod 12$ (coset structure), not the complete discrete log. This shows: - Manifold discovery can be “partial”—discover quotient group first, then (possibly) discover complete group - This is consistent with the two-stage model: local manifold discovery \rightarrow global gluing

4. Dimension Lower Bound and Task Complexity

Modular addition critical point 8-16 dim, modular multiplication 16-32 dim. This matches intuition: multiplicative group structure is more complex (has non-trivial subgroups), needs more dimensions to encode.

5. Topological Possession: Topology Competition Under Capacity Constraints

The nested Grokking experiment revealed a new phenomenon: when model capacity is insufficient to simultaneously maintain two topological levels, instead of coexistence, **topological possession**

occurs—the later-learned structure replaces the earlier-learned one, while output remains unchanged throughout. This shows: - Representation space structure can undergo complete reorganization while output remains perfectly unchanged - **test_acc=100% does not mean the model’s internal understanding is stable**—beneath the same accuracy, entirely different topological organizations may exist - Manifold discovery is not a unidirectional “bad to good” process, but dynamic competition among multiple topological candidate solutions

6. Capacity-Regularization Matching Principle

The failure of the scaling experiment (4-layer, 256-dim, test_acc=51.75%) overturns the naive intuition that “bigger is better.” The large model momentarily touched perfect topology at step 360K then immediately collapsed, showing the target solution **exists** in parameter space, but the optimization landscape is too flat for the optimizer to find a stable path. This suggests a **matching principle**: capacity and regularization pressure must be jointly tuned—increasing capacity alone without increasing constraints is like giving a bigger house with no furniture; the structure is actually less organized than what was forced out of the smaller house.

7.4 Limitations and Future Directions

Limitations of this paper: 1. **Limited task scope**: Only validated on modular addition and multiplication, unclear if applicable to other Grokking tasks (like permutation groups, polynomial evaluation) 2. **Two-stage model pending verification**: Will modular multiplication’s “global gluing” occur with longer training? (Nested Grokking experiments show: extended training + increased WD does not lead to global gluing, but rather topological possession, and at 5M steps the structure is destroyed entirely) 3. **64-dimension anomaly pending explanation**: Currently only intuitive explanation, lacking rigorous theory 4. **Quantitative laws of capacity-regularization matching**: How much WD does a large model need to force structure? Does a scaling law like $WD \propto \sqrt{\text{parameters}}$ exist?

Future directions: 1. Validate two-stage model and topological possession phenomenon on more tasks 2. Explore whether Grokking can be controlled by manipulating representation space topology 3. Study the mathematical mechanism of 64-dimension anomaly 4. **Joint capacity-regularization sweep**: Systematically search the 2D phase diagram of model size and WD, looking for the critical curve from “possession” to “coexistence” 5. **Switch to gcd=1 tasks** (e.g., mod 91 = 7×13), verify whether two-level structures are more easily discovered simultaneously when naturally orthogonal 6. **Reproducibility of topological flashing**: Is the large model’s momentary perfect topology at 360K/380K steps coincidental or a necessary waypoint? Can it be reproduced across multiple seeds?

8. Conclusion

Grokking is not an anomaly, but a window into deep learning—it reveals the essence of generalization.

The manifold discovery hypothesis proposed in this paper provides a unified framework from the internal perspective:

- **Memorization = jagged curve** (high-dimensional, sample-wise encoding)
- **Generalization = smooth manifold** (low-dimensional, structure-wise encoding)
- **Grokking = phase transition in multi-stable system** (possibly accompanied by critical state oscillation)

- **Weight decay = centripetal force that destabilizes jagged curves**

We validated this hypothesis on two experimental groups:

Metric	Mod Addition	Mod Multiplication
Dimension change	$78 \rightarrow 8$	$89 \rightarrow 11$
Topological change	β_0 500 \rightarrow 6, β_1 504 \rightarrow 0	β_0 unchanged, β_1 870 \rightarrow 179
Bottleneck critical point	8-16 dim	16-32 dim
Manifold structure	97 clusters	12 cosets (purity 99.4%)
Adjacency score	0% (no ring topology)	100% (perfect \mathbb{Z}_{12} ring)
Multi-seed success rate	67% (2/3)	67% (2/3)
Oscillation count	16.3 ± 2.9	14.3 ± 2.1

Based on experimental findings, we revised the hypothesis into a **two-stage model**: 1. **Local manifold discovery**: Intra-component structuring (dimension drops, β_1 drops) 2. **Global gluing**: Inter-component alignment (β_0 drops)

Modular addition completed both stages, modular multiplication only completed the first stage—this explains why modular multiplication learned quotient group structure rather than the complete multiplicative group.

Nested Grokking experiments further revealed: - **Topological possession**: When capacity is insufficient to encode multiple structural levels simultaneously, the model completely replaces its internal topology while maintaining test_acc=100% - **WD non-monotonicity**: Regularization first forces structure, then destroys it; an optimal training length exists - **Inner layer walks gcd**: The model autonomously discovers $\gcd(12, 8) = 4$ as the shortest path for simultaneously encoding both layers of information - **Capacity-regularization matching principle**: Simply increasing capacity (8x parameters) not only failed to achieve topological coexistence, but caused Grok failure entirely (test_acc=51.75%). Capacity and regularization pressure must be jointly tuned

In one sentence: high-dimensional curve \rightarrow oscillation \rightarrow low-dimensional surface (success rate ~67%, possibly in two steps; under capacity constraints, multi-level structures undergo topological competition).

References

1. Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. arXiv:2201.02177.
2. Liu, Z., Kitouni, O., Nolte, N., Michaud, E. J., Tegmark, M., & Williams, M. (2022). Towards Understanding Grokking: An Effective Theory of Representation Learning. NeurIPS 2022. arXiv:2205.10343.
3. Liu, Z., Michaud, E. J., & Tegmark, M. (2023). Omnigrok: Grokking Beyond Algorithmic Data. ICLR 2023. arXiv:2210.01117.

4. Prieto, L., Barsbey, M., Mediano, P. A. M., & Birdal, T. (2025). Grokking at the Edge of Numerical Stability. ICLR 2025. arXiv:2501.04697.
5. Kumar, T., Bordelon, B., Gershman, S. J., & Pehlevan, C. (2024). Grokking as the Transition from Lazy to Rich Training Dynamics. ICLR 2024. arXiv:2310.06110.
6. Varma, V., Shah, R., Kenton, Z., Kramár, J., & Kumar, R. (2023). Explaining Grokking Through Circuit Efficiency. arXiv:2309.02390.
7. Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress Measures for Grokking via Mechanistic Interpretability. ICLR 2023 (Oral). arXiv:2301.05217.
8. Facco, E., d’Errico, M., Rodriguez, A., & Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. Scientific Reports, 7(1), 12140.
9. Carlsson, G. (2009). Topology and data. Bulletin of the American Mathematical Society, 46(2), 255-308.

“Existing research primarily asks ‘under what conditions does Grokking occur’; this paper attempts to ask ‘what is Grokking.’ The former is an engineering question; the latter is an ontological question.”