

Contents

Grokking as Manifold Discovery: A Geometric Reinterpretation of Delayed Generalization	2
1. Introduction: Why Grokking Matters	2
2. Review of Existing Theories	3
2.1 Goldilocks Zone Theory (Liu et al. 2022)	3
2.2 Softmax Collapse Theory (Prieto et al. 2025)	3
2.3 Lazy → Rich Transition Theory (Kumar et al. 2024)	3
2.4 Weight Efficiency Hypothesis (Varma et al. 2023)	4
2.5 Mechanistic Interpretability Perspective (Nanda et al. 2023)	4
2.6 Common Blind Spot of Existing Theories	4
3. Unified Framework: The Manifold Discovery Hypothesis	5
3.1 Geometric Interpretation of Memorization vs. Generalization	5
3.2 Geometric Role of Weight Decay	6
3.3 Geometric Interpretation of Softmax Collapse	6
3.4 Geometric Interpretation of Lazy → Rich	6
3.5 Geometric Interpretation of Weight Efficiency	7
3.6 A Set of Formalizable Mathematical Statements (Turning “Metaphors” into Provable Propositions)	7
3.7 A Formulation Closer to Real LLMs (GPT-4 Class Systems): Continuous Approximation + Spectral/Low-Complexity Bias	9
4. Reinterpreting Existing Findings	10
4.1 Why Weight Decay Being Too Large/Small Both Fail	10
4.2 Why Data Amount Affects Grokking	10
4.3 Why Over-parameterized Models Are More Prone to Grokking	10
4.4 Why Grokking Is Sudden (and Why It Oscillates)	11
5. Testable Predictions	11
5.1 Intrinsic Dimension Discontinuity	11
5.2 Topological Structure of Representations	12
5.3 Attention Entropy Dynamics	12
5.4 Effect of Forced Rank Constraints	12
6. Experimental Validation	13
6.1 Experimental Configuration	13
6.2 Experiment Group 1: Modular Addition Results	14
6.3 Experiment Group 2: Modular Multiplication Results	16
6.4 Coset Structure Verification: $12 \text{ Clusters} = k \bmod 12$	17
6.5 Hypothesis Revision: Two-Stage Model	17
6.6 Adjacency Analysis: Differences in Topology Preservation	18
6.7 Multi-Seed Stability Verification	18
6.8 Cross-Operation Comparison of Two Experimental Groups	19
7. Discussion	19
7.1 Methodological Limitations of Existing Research	19
7.2 Value of the Internal Perspective	20
7.3 Theoretical Significance of Experimental Findings	20
7.4 Limitations and Future Directions	20
8. Conclusion	21
References	22