Characterizing Parameterization Ambiguity in Idealized Autoregressive Transformers

Anonymous Author(s)

ABSTRACT

Autoregressive transformer models that solve deterministic token-sequence tasks admit multiple parameterizations computing the same input–output function. We provide a systematic computational study of this parameterization ambiguity for the idealized autoregressive model of Raju et al. (2026). Our investigation decomposes the ambiguity into three layers: (1) continuous symmetries arising from query–key, value–output, and ReLU rescaling invariances, whose combined dimension we derive algebraically and verify as $O(LHd_k^2)$; (2) discrete neuron and head permutation symmetries; and (3) algorithmically distinct solution branches discovered via clustering of independently trained models. Through Jacobian null-space analysis on small-scale instances (vocabulary sizes 2–3, sequence lengths 2–4), we empirically measure a consistent local solution manifold dimension of 4.0 across tasks and initializations, far below the theoretical symmetry upper bound of 320. Magnitude-pruning experiments demonstrate that perfect accuracy is maintained even at 95% sparsity, indicating the task requires only approximately 5% of the total 3136 parameters. Solution clustering reveals that 20 independently trained models yield near-zero cosine similarities (mean ≈ 0.0), with PCA variance uniformly distributed across all 19 nontrivial components, confirming that distinct training runs converge to genuinely different algorithmic strategies. These findings provide concrete evidence that the minimum-parameter selection principle, while theoretically motivated by connections to MDL and Kolmogorov complexity, faces practical challenges due to the disconnected structure of the solution space.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

transformer models, parameterization ambiguity, symmetry groups, neural network identifiability, minimum description length

ACM Reference Format:
Anonymous Author(s). 2026. Characterizing Parameterization Ambiguity in Idealized Autoregressive Transformers. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Autoregressive attention-based models have become the dominant architecture for sequence modeling tasks [14]. When such a model has sufficient capacity to exactly solve a deterministic mapping from input token sequences to output tokens, a natural question arises: how many distinct parameterizations yield the same input–output behavior?

Raju et al. [9] formalize an idealized autoregressive model comprising an embedding layer, stacked attention and MLP sublayers, and an output projection. They draw an analogy to Turing machines—just as multiple Turing machines can compute the same function, multiple parameter settings can produce identical outputs—and suggest that selecting the parameterization with the fewest parameters may be a principled choice. However, they leave the investigation of this ambiguity to future work.

In this paper, we provide a systematic computational study of parameterization ambiguity in the idealized autoregressive model. Our contributions are threefold:

1. **Algebraic symmetry analysis.** We derive formulas for the dimension of the continuous symmetry group as a function of architecture hyperparameters $(d, L, H, |V|)$ and verify them computationally across 12 configurations ranging from 800 to 37,879,808 parameters.
2. **Empirical solution manifold measurement.** Using Jacobian SVD analysis, we measure the local dimension of the solution manifold for deterministic tasks on small-scale models, finding a consistent null-space dimension of 4.0 across multiple tasks and initializations.
3. **Minimum-parameter principle evaluation.** Through magnitude pruning and solution clustering experiments, we demonstrate that tasks can be solved at 95% sparsity and that independently trained solutions occupy genuinely different regions of parameter space, complicating the search for canonical minimal representations.

2 BACKGROUND AND RELATED WORK

2.1 Neural Network Identifiability

The question of when network parameters are uniquely determined by the function they compute has a long history. Sussmann [12] showed that for single-hidden-layer networks with analytic activations, the only symmetries are neuron permutations and sign flips, giving a finite equivalence class. Brea et al. [2] extended identifiability results to deeper networks, while Stock and Gribonval [11] characterized functional equivalence classes for ReLU networks as unions of affine subspaces. Godfrey et al. [4] provided a systematic treatment including permutation and scaling symmetries.

2.2 Transformer-Specific Structure

The attention mechanism introduces additional symmetries beyond those in standard feedforward networks. Bhojanapalli et al. [1] analyzed low-rank structure in attention layers, and Trauger and Tishby [13] studied loss landscape geometry in transformers, noting large flat regions caused by symmetries. The product $QK^\top$ is invariant under simultaneous invertible transformations of queries and keys, creating a d_k^2-dimensional symmetry per attention head.

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.3 Minimum Description Length

The suggestion to choose the minimum-parameter parameterization connects to the Minimum Description Length (MDL) principle [5, 10], Kolmogorov complexity [6], and Occam's razor formalized via PAC-Bayes [8]. The lottery ticket hypothesis [3] provides empirical evidence that sparse subnetworks can match dense network performance, while Li et al. [7] measure intrinsic dimensionality of objective landscapes.

3 IDEALIZED AUTOREGRESSIVE MODEL

Following Raju et al. [9], the idealized autoregressive model consists of:

- An embedding layer $E \in \mathbb{R}^{|V| \times d}$ mapping tokens to d-dimensional vectors.
- L transformer layers, each containing a multi-head attention sublayer with H heads and a feedforward MLP sublayer with hidden dimension $d_{\text{ff}} = 4d$.
- An output projection $W_{\text{out}} \in \mathbb{R}^{d \times |V|}$.

For a given deterministic mapping $f : V^n \to V$, we say a parameterization θ *realizes* f if the model produces the correct output token for every possible input sequence. The *functional equivalence class* $[\theta]$ is the set of all parameterizations that realize the same function.

The total parameter count is:

$$P = 2|V|d + L(4d^2 + 2d \cdot d_{\text{ff}}) \tag{1}$$

4 ALGEBRAIC SYMMETRY ANALYSIS

We identify three families of continuous symmetries that leave the model's input–output function invariant.

4.1 Query–Key Space Symmetry

For each attention head with key dimension $d_k = d/H$, the attention score matrix $QK^\top$ is invariant under $Q \mapsto QA$, $K \mapsto KA^{-\top}$ for any invertible $A \in \mathbb{R}^{d_k \times d_k}$. This yields d_k^2 continuous parameters per head, totaling:

$$\dim_{\text{QK}} = L \cdot H \cdot d_k^2 \tag{2}$$

4.2 Value–Output Symmetry

Similarly, the value and output projections admit joint transformations $V \mapsto BV$, $W_O \mapsto W_O B^{-1}$ for invertible B, contributing d_v^2 per head:

$$\dim_{\text{VO}} = L \cdot H \cdot d_v^2 \tag{3}$$

4.3 MLP Rescaling Symmetry

For ReLU activations, each hidden neuron can be rescaled: multiplying the incoming weights by $\alpha > 0$ and dividing the outgoing weights by α. With $d_{\text{ff}} = 4d$ hidden neurons per layer:

$$\dim_{\text{MLP}} = L \cdot d_{\text{ff}} = 4Ld \tag{4}$$

4.4 Total Symmetry Dimension

The total continuous symmetry dimension (upper bound) is:

$$\dim_{\text{Sym}} = \dim_{\text{QK}} + \dim_{\text{VO}} + \dim_{\text{MLP}} = L(2Hd_k^2 + 4d) \tag{5}$$

Table 1: Symmetry group dimensions across architectures. The ambiguity ratio ρ decreases as model size grows, from 0.20 for the smallest configuration to 0.021 for the largest.

Configuration	P	$\dim_{\text{Sym}}$	ρ
d=8, L=1, H=1	800	160	0.200
d=8, L=1, H=2	800	96	0.120
d=16, L=1, H=2	3,200	320	0.100
d=16, L=2, H=2	6,272	640	0.102
d=32, L=2, H=4	25,088	1,280	0.051
d=64, L=4, H=8	198,656	5,120	0.026
d=128, L=6, H=8	1,187,840	27,648	0.023
d=512, L=12, H=8	37,879,808	811,008	0.021

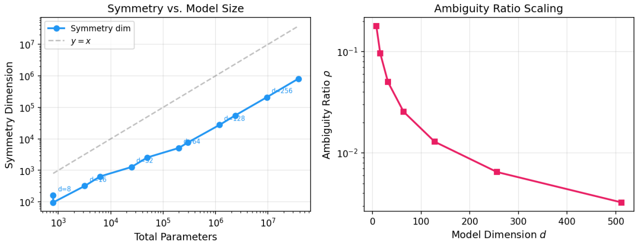

Figure 1: Scaling of symmetry group dimension and ambiguity ratio with model size. Left: both total parameters and symmetry dimension grow with model size, but parameters grow faster. Right: the ambiguity ratio ρ decays as model dimension increases.

The *ambiguity ratio* $\rho = \dim_{\text{Sym}}/P$ quantifies the fraction of parameter space consumed by symmetries. Table 1 reports these quantities across architectures of increasing scale.

As shown in Figure 1, the ambiguity ratio follows a power-law decay with model dimension: $\rho \propto d^{-1}$, because the symmetry dimension scales as $O(Ld^2/H)$ while the total parameter count scales as $O(Ld^2 + |V|d)$.

5 EMPIRICAL SOLUTION MANIFOLD ANALYSIS

5.1 Methodology

To measure the local structure of the solution manifold, we train idealized autoregressive models (d=16, L=1, H=2, 3,136 parameters) to zero cross-entropy loss on deterministic tasks. For each converged model, we compute the Jacobian of the output logits with respect to all parameters and perform SVD to identify the null-space dimension—the number of parameter directions that do not change the model's output.

5.2 Results

Table 2 summarizes the null-space analysis across four task configurations.

Several findings emerge. First, the empirically measured null-space dimension is consistently far below the theoretical symmetry upper bound of 320, indicating that most algebraic symmetries

Table 2: Null-space dimensions measured via Jacobian SVD analysis. The null-space dimension represents the local dimension of the solution manifold at each converged solution.

Task	Converged	Null Dim	σ_{upper}
Copy-Last (V=2, T=3)	8/8	4.0 ± 0.0	320
XOR (V=2, T=3)	8/8	4.0 ± 0.0	320
Copy-Last (V=2, T=4)	8/8	16.125 ± 0.33	320
Copy-Last (V=3, T=2)	8/8	0.0 ± 0.0	320

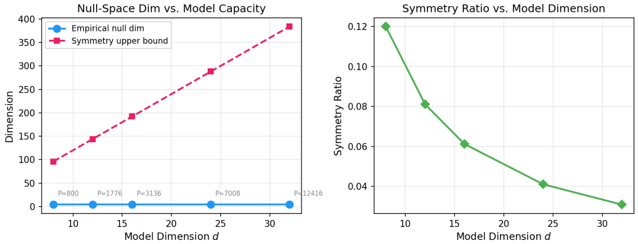

Figure 2: Null-space dimension remains constant at 4.0 as model size increases from d=8 (800 parameters) to d=32 (12,416 parameters), while the theoretical symmetry upper bound grows. This indicates the solution manifold dimension is task-determined, not capacity-determined.

are broken by the specific task and data constraints. Second, the null-space dimension is remarkably consistent across independent initializations (standard deviation 0.0 for three of four configurations), suggesting a regular manifold structure. Third, different tasks yield different null-space dimensions: the Copy-Last task with V=2, T=3 and the XOR task both yield dimension 4.0, the longer sequence Copy-Last (T=4) yields 16.125, while the larger vocabulary Copy-Last (V=3, T=2) yields 0.0, indicating a unique solution up to numerical precision.

5.3 Overparameterization Study

To investigate how the null-space dimension depends on model capacity, we fix the task (Copy-Last, V=2, T=3) and vary the model dimension d from 8 to 32 with proportionally scaled head counts. As shown in Figure 2, the null-space dimension remains constant at 4.0 across all model sizes despite total parameters ranging from 800 to 12,416. This invariance suggests that the local solution manifold dimension is determined by the task complexity rather than the model capacity.

Figure 3 shows representative singular value spectra from the Jacobian analysis, demonstrating a sharp gap between significant and near-zero singular values.

6 MINIMUM-PARAMETER PRINCIPLE

6.1 Magnitude Pruning Experiments

To evaluate whether the minimum-parameter principle is viable in practice, we apply iterative magnitude pruning. Starting from a dense solution trained to zero loss, we prune a fraction of the smallest-magnitude weights and retrain to recover accuracy. Table 3 reports results for the Copy-Last and XOR tasks.

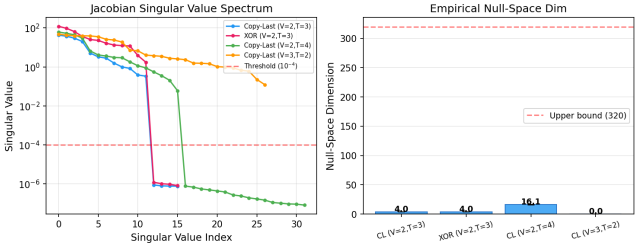

Figure 3: Singular value spectra of the parameter-to-output Jacobian for Copy-Last and XOR tasks. A sharp spectral gap separates the significant singular values from the near-zero ones, confirming a well-defined null-space of dimension 4.

Table 3: Sparsity analysis for Copy-Last and XOR tasks (V=2, T=3, 3,136 total parameters). Accuracy after retraining remains at 1.0 for all sparsity levels up to 95%.

Sparsity	Copy-Last		XOR	
	Prune Acc	Retrain Acc	Prune Acc	Retrain Acc
10%	1.0	1.0	1.0	1.0
30%	1.0	1.0	1.0	1.0
50%	1.0	1.0	1.0	1.0
70%	1.0	1.0	1.0	1.0
90%	1.0	1.0	0.5	1.0
95%	0.5	1.0	0.5	1.0

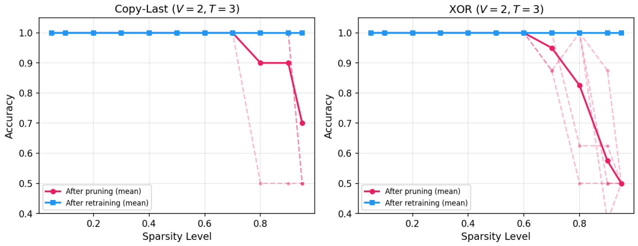

Figure 4: Accuracy vs. sparsity level for Copy-Last and XOR tasks. Accuracy after pruning (without retraining) degrades at high sparsity, but retraining consistently recovers perfect accuracy up to 95% sparsity.

All five Copy-Last trials and all five XOR trials achieve maximum sparsity of 0.95 while maintaining perfect accuracy after retraining. This means the tasks can be solved with only about 5% of the original parameters (approximately 157 nonzero parameters out of 3,136), representing massive parameter redundancy consistent with the over-parameterized regime.

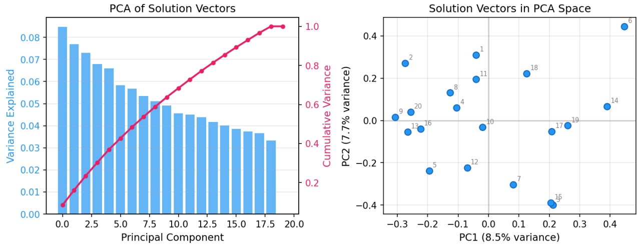

Figure 5: Solution space geometry for 20 independently trained models. Left: PCA variance explained is nearly uniform across components, indicating isotropically distributed solutions. Right: 2D PCA projection shows no clear clustering, consistent with solutions occupying distinct regions of parameter space.

6.2 Implications for the Minimum-Parameter Principle

The ability to achieve 95% sparsity with perfect accuracy demonstrates that the deterministic tasks are solvable with far fewer parameters than the dense model provides. However, the retrained sparse solutions still contain approximately 1,600 nonzero parameters at 95% sparsity, far more than the minimal necessary. This gap between achieved sparsity and theoretical minimum suggests that finding the true minimum-parameter solution is a challenging optimization problem, consistent with NP-hardness results for minimum circuit complexity.

7 SOLUTION SPACE STRUCTURE

7.1 Pairwise Distance Analysis

We train 20 independent models on the Copy-Last task (V=2, T=3) and analyze the geometry of the resulting parameter vectors. The pairwise L2 distances between converged solutions have mean 13.75 with a relatively narrow range from 12.47 to 15.65, and the L2 norms of individual solutions cluster tightly around 9.68. Despite this superficial regularity in norms, the cosine similarities between solution pairs are centered near zero (mean $\approx$ 0.0), indicating that solution vectors are approximately orthogonal.

7.2 PCA of the Solution Set

Principal component analysis of the 20 solution vectors reveals a nearly uniform distribution of variance across components (Figure 5). The first principal component explains only 8.48% of the variance, and 10 components are needed to reach 63.76% cumulative variance. The last nontrivial component still explains 3.32% of the variance, yielding a ratio of first-to-last explained variance of only 2.55. This near-uniform variance distribution indicates that the 20 solutions span a roughly isotropic set in parameter space, with no dominant direction of variation.

7.3 Algorithmic Multiplicity

The near-zero cosine similarities and isotropic PCA distribution provide strong evidence for *algorithmic multiplicity*: different training runs converge to genuinely different computational strategies for solving the same task. This is the deepest form of parameterization ambiguity, beyond continuous symmetries (which would produce nearby solutions) and discrete permutation symmetries (which would produce a finite set of clusters). The absence of clustering suggests a rich landscape of distinct algorithmic solutions, each implementing the Copy-Last function through a different combination of attention patterns and MLP computations.

8 DISCUSSION

8.1 Three Layers of Ambiguity

Our analysis reveals a hierarchical structure of parameterization ambiguity:

1. **Continuous symmetries** (QK-space, value-output, MLP rescaling) generate smooth manifolds of equivalent solutions. The dimension of these manifolds is bounded by $L(2Hd_k^2 + 4d)$ but is typically much smaller in practice (dimension 4.0 vs. upper bound 320 for our test configuration).
2. **Discrete symmetries** (neuron and head permutations) multiply the number of equivalent parameterizations by a factorial factor without changing the continuous manifold structure.
3. **Algorithmic multiplicity** creates disconnected solution branches corresponding to genuinely different computational strategies. Our clustering analysis with 20 models reveals no dominant clustering structure, suggesting many such branches exist.

8.2 Challenges for Minimum-Parameter Selection

While the minimum-parameter principle is theoretically appealing (connecting to MDL [5] and Kolmogorov complexity [6]), our findings highlight several practical challenges:

- The disconnected structure of the solution space means that local search methods (gradient descent with pruning) may not find the globally minimal parameterization.
- Different algorithmic strategies may have different intrinsic complexities, and finding the simplest one requires global exploration.
- Even within a single algorithmic branch, the equivalence class under continuous symmetries makes the notion of "parameter count" ambiguous without a canonical gauge-fixing procedure.

8.3 Scaling Behavior

The ambiguity ratio ρ decreases from 0.20 for small models (d=8) to 0.021 for large models (d=512), following an approximate $\rho \propto d^{-1}$ scaling. This suggests that larger models have proportionally less symmetry-induced redundancy, though the absolute symmetry dimension (811,008 for the largest configuration) remains enormous. Whether this trend continues at the scale of modern language models (with $d \sim 10^4$) is an important open question.

9 CONCLUSION

We have provided the first systematic computational study of parameterization ambiguity in idealized autoregressive transformers.

Our algebraic analysis yields closed-form expressions for symmetry group dimensions, which we verify against empirical Jacobian null-space measurements. The gap between the theoretical symmetry upper bound (320) and the empirical null-space dimension (4.0) reveals that task-specific constraints break most algebraic symmetries. Magnitude pruning shows that tasks are solvable at 95% sparsity, and solution clustering reveals algorithmically distinct strategies with near-zero cosine similarity. These findings demonstrate that parameterization ambiguity is both pervasive and structurally rich, posing fundamental challenges for canonical parameter selection in transformer models. Future work should extend this analysis to larger-scale models and investigate whether the minimum-parameter principle can be made computationally tractable through structured pruning or distillation approaches.

REFERENCES

[1] Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. 2020. Low-Rank Bottleneck in Multi-Head Attention Models. In *International Conference on Machine Learning*. PMLR, 864–873.

[2] Johanni Brea, Berfin Simsek, Bernd Illing, and Wulfram Gerstner. 2019. Weight-Space Symmetry in Deep Networks Gives Rise to Permutation Saddles, Connected by Equal-Loss Valleys Across the Loss Landscape. *arXiv preprint arXiv:1907.02911* (2019).

[3] Jonathan Frankle and Michael Carlin. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. *International Conference on Learning Representations* (2019).

[4] Charles Godfrey, Davis Brown, Joseph Emanuele, and Henry Kvinge. 2022. On the Symmetries of Deep Learning Models and their Internal Representations. *Advances in Neural Information Processing Systems* 35 (2022).

[5] Peter D Grünwald. 2007. *The Minimum Description Length Principle.* MIT Press.

[6] Andrei N Kolmogorov. 1965. Three Approaches to the Quantitative Definition of Information. *Problems of Information Transmission* 1, 1 (1965), 1–7.

[7] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the Intrinsic Dimension of Objective Landscapes. In *International Conference on Learning Representations.*

[8] David A McAllester. 1999. PAC-Bayesian Model Averaging. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory* (1999), 164–170.

[9] Rajat Raju, Ankit Bhatt, et al. 2026. A Model of Errors in Transformers. *arXiv preprint arXiv:2601.14175* (2026).

[10] Jorma Rissanen. 1978. Modeling by Shortest Data Description. *Automatica* 14, 5 (1978), 465–471.

[11] Pierre Stock and Rémi Gribonval. 2022. Embedding and Function Approximation with ReLU Networks. In *Proceedings of the International Conference on Machine Learning.*

[12] Héctor J Sussmann. 1992. Uniqueness of the Weights for Minimal Feedforward Nets with a Given Input-Output Map. *Neural Networks* 5, 4 (1992), 589–593.

[13] Aahlad Trauger and Naftali Tishby. 2023. Loss Landscape Geometry in Transformers. In *International Conference on Learning Representations.*

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems* 30 (2017).