Bispectral Invariants for Transformers: An Operator-Algebraic Approach

Abstract

Modern Transformer models exhibit massive parameter redundancy, with millions of distinct configurations yielding identical functions. We provide the first complete characterization of this phenomenon through the maximal gauge group $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$. Our approach combines operator-algebraic methods with harmonic analysis to develop complete computational invariants for representation equivalence. We formalize Transformer layers as modules over C*-algebras, enabling rigorous analysis via Morita theory and Fredholm indices. The centerpiece of our framework is the G-bispectrum, which, after canonical gauge-fixing to eliminate continuous degrees of freedom, provides complete invariants for the residual permutation symmetry S_h. We introduce a selective variant achieving $O(h)$ complexity for permutation group discrimination after canonicalization. Comprehensive experiments validate our theory: gauge transformations preserve outputs to machine precision (relative error bounded by $15\varepsilon_{\mathrm{mach}}$), the bispectrum achieves 100% discrimination between non-equivalent canonicalized models, and the selective variant provides 42.7× speedup for permutation group analysis. These results establish foundational tools for model comparison, optimization analysis, and understanding the true complexity of Transformer representations.

1. Introduction

The remarkable success of Transformer models [26; 4; 1] has revolutionized natural language processing and beyond. However, these models exhibit massive parameter redundancy where many distinct parameter configurations yield identical input-output functions. This redundancy has profound implications for optimization [16; 24], model comparison [14; 5], interpretability [7; 17], and compression [8; 20].

Consider a standard 12-head attention layer with dimension 768. The parameter space has over 9 million dimensions, yet the functional degrees of freedom are far fewer due to internal symmetries. Which transformations preserve the model function? Can we efficiently determine if two models are equivalent? What are the true degrees of freedom? These questions motivate our theoretical investigation.

1.1. Motivating Example

Consider the attention computation for head i in a multi-head attention layer. The queries, keys, and values are computed as $Q_i = XW_Q^{(i)}$, $K_i = XW_K^{(i)}$, and $V_i = XW_V^{(i)}$ respectively. The attention mechanism then computes

$$\mathrm{Attn}_i = \mathrm{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i.$$

A key observation is that the transformation $(W_Q^{(i)}, W_K^{(i)}) \mapsto (W_Q^{(i)} A, W_K^{(i)} (A^{-1})^T)$ for any invertible matrix $A \in \mathrm{GL}(d_k)$ preserves the product $Q_i K_i^T$ and thus the attention pattern. Similarly, the transformation $(W_V^{(i)}, W_O^{(i)}) \mapsto (W_V^{(i)} C, C^{-1} W_O^{(i)})$ for $C \in \mathrm{GL}(d_v)$ preserves the composed value-output mapping. These symmetries create vast equivalence classes in parameter space, fundamentally affecting optimization landscapes and model comparison.

1.2. Contributions

This paper provides the first complete algebraic characterization of Transformer symmetries through the following contributions:

Exact Symmetry Group (Section 3): We prove that $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$ is the maximal gauge group for standard multi-head attention, characterizing all parameter transformations that preserve the model function. The proof (Appendix A) uses differential-geometric methods to establish both necessity and sufficiency.

Operator-Algebraic Framework (Section 4): We formulate Transformer computations as modules over C*-algebras, enabling analysis via Morita equivalence and providing a categorical perspective on representation equivalence. This framework connects to broader mathematical structures in operator theory.

Information-Theoretic Characterization (Section 5): We introduce the Fredholm index as a gauge-invariant measure of information flow, revealing the characteristic expansion-compression pattern in Transformer blocks that maintains zero net information loss.

Complete Computational Invariants (Section 6): Building on the pioneering work of Sanborn and Miolane [19] who demonstrated the power of bispectral methods for analyzing neural network representations, we develop the G-bispectrum as a complete invariant for gauge orbits. We introduce a selective variant achieving O(h) complexity for permutation group discrimination after canonicalization.

Comprehensive Validation (Section 7): Extensive experiments confirm machine-precision invariance, perfect discrimination accuracy, and significant computational speedups with the selective variant, validating all theoretical predictions.

2. Background and Preliminaries

2.1. Multi-Head Attention Architecture

The standard multi-head attention mechanism [26] operates with h attention heads on input sequences $X \in \mathbb{R}^{n \times d_{\text{model}}}$. Each head $i \in \{1, \ldots, h\}$ computes queries $Q_i = XW_Q^{(i)}$, keys $K_i = XW_K^{(i)}$, and values $V_i = XW_V^{(i)}$ where $W_Q^{(i)}, W_K^{(i)} \in \mathbb{R}^{d_{\text{model}} \times d_k}$ and $W_V^{(i)} \in \mathbb{R}^{d_{\text{model}} \times d_v}$.

The attention mechanism computes weighted combinations of values based on query-key similarities. Each head produces output

$$\text{head}_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i,$$

and the multi-head attention output combines all heads through concatenation and projection:

$$\text{MHA}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W_O$$

where $W_O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$. Here W_O denotes the single output projection matrix, decomposed into blocks $W_O^{(i)} \in \mathbb{R}^{d_v \times d_{\text{model}}}$ when analyzing per-head contributions.

2.2. Group Theory Preliminaries

A group action of G on a set X is a map $G \times X \to X$ satisfying identity and compatibility conditions. The orbit of $x \in X$ under the group action is $G \cdot x = \{g \cdot x : g \in G\}$, and an invariant is a function that remains constant on orbits.

The general linear group $\mathrm{GL}(n)$ consists of all invertible $n \times n$ matrices, capturing linear transformations that preserve vector space structure. The symmetric group S_n consists of all permutations of n elements. A semi-direct product $G \rtimes H$ combines groups where H acts on G by automorphisms, capturing both internal symmetries and their interactions.

2.3. Harmonic Analysis Background

Harmonic analysis on groups provides tools for decomposing functions into fundamental frequency components. For a group G, the Fourier transform decomposes functions $f : G \to \mathbb{C}$ into irreducible representations, revealing the underlying structure.

The bispectrum, introduced by Kakarala [11] and further developed by Kondor [12], captures third-order correlations and provides complete invariants for group actions under appropriate conditions. Sanborn and Miolane's recent work [19] has been particularly influential in demonstrating how bispectral analysis can reveal hidden structure in neural network representations. Their key insight that neural networks naturally exhibit group-theoretic structure that can be captured through harmonic analysis directly inspired our approach to characterizing Transformer gauge symmetries through the G-bispectrum.

3. Maximal Gauge Group of Transformers

We now establish the complete symmetry structure of multi-head attention through a rigorous characterization of parameter transformations that preserve model functionality.

Definition 1 (Parameter Space and Model Function) *The parameter space of multi-head attention is*

$$\Theta = \prod_{i=1}^{h} (\mathbb{R}^{d_{model} \times d_k} \times \mathbb{R}^{d_{model} \times d_k} \times \mathbb{R}^{d_{model} \times d_v}) \times \mathbb{R}^{hd_v \times d_{model}}$$

representing all weight matrices $(W_Q^{(i)}, W_K^{(i)}, W_V^{(i)})$ *for each head and the output projection* W_O. *The model function is* $F_\theta : \mathbb{R}^{n \times d_{model}} \to \mathbb{R}^{n \times d_{model}}$.

Definition 2 (Gauge Transformation) *A gauge transformation is a differentiable map* $g : \Theta \to \Theta$ *such that* $F_{g(\theta)} = F_\theta$ *for all* $\theta \in \Theta$. *The set of all gauge transformations forms a group under composition, called the gauge group.*

Theorem 3 (Maximal Gauge Group) *For standard multi-head attention with separate query, key, value projections per head, single output projection applied to concatenated heads, full column rank projection matrices, LayerNorm applied after output projection, and no positional encodings or architectural constraints, the maximal gauge group is:*

$$G_{\max} = \left((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h \right) \rtimes S_h$$

The group acts on parameters by transforming each head's projections with invertible matrices while allowing permutations among heads. The proof, detailed in Appendix A, establishes this through analysis of infinitesimal symmetries (Lemma 20 and Proposition 21) and the necessary factorization structure (Theorem 23).

Corollary 4 (Dimension of Gauge Orbits) *Each* $G_{\max}$*-orbit has continuous dimension* $h(d_k^2 + d_v^2)$, *with additionally* $h!$ *discrete representatives from head permutations* S_h.

4. Operator-Algebraic Framework

We develop an operator-algebraic formulation that provides powerful mathematical tools for analyzing Transformer representations and their equivalences.

4.1. C*-Algebra Formulation

Definition 5 (Transformer C*-Algebra) *Let $\mathcal{H} = \ell^2(\mathbb{N}) \otimes \mathbb{C}^{d_{model}}$ be the Hilbert space of sequences. The Transformer C*-algebra $\mathcal{A} \subset B(\mathcal{H})$ is the norm-closed algebra generated by linear attention operators $T_{attn} : \mathcal{H} \to \mathcal{H}$ and feedforward operators $T_{ffn} : \mathcal{H} \to \mathcal{H}$. LayerNorm is treated as an external normalization operation, not an element of $\mathcal{A}$.*

Proposition 6 (Gauge Action on $\mathcal{A}$) *The gauge group $G_{\max}$ acts on $\mathcal{A}$ by algebra automorphisms: $\alpha_g(T) = S_g T S_g^{-1}$ for $g \in G_{\max}$ and $T \in \mathcal{A}$, where S_g is a bounded invertible operator implementing g on $\mathcal{H}$. For elements $k \in K \subset G_{\max}$ in the maximal compact subgroup $K = (O(d_k))^h \times (O(d_v))^h \rtimes S_h$, the action is by *-automorphisms via unitary operators U_k.*

4.2. Morita Equivalence and Representation Categories

Definition 7 (Transformer Modules) *A parameter configuration $\theta \in \Theta$ induces a right $\mathcal{A}$-module $\mathcal{M}_\theta$ with action given by layer composition. The module structure encodes how layers transform representations.*

Theorem 8 (Morita Equivalence Characterization) *Two parameter configurations $\theta, \theta' \in \Theta$ satisfy $\theta' = g \cdot \theta$ for some $g \in G_{\max}$ if and only if the induced modules $\mathcal{M}_\theta$ and $\mathcal{M}_{\theta'}$ are Morita equivalent as $\mathcal{A}$-modules.*

This theorem, proven in Appendix D, provides a categorical characterization where gauge-equivalent parameters yield equivalent representation categories. The Morita equivalence framework reveals that seemingly different parameter configurations can implement identical computational structures.

4.3. Von Neumann Algebra Extension

For analyzing infinite-depth limits and asymptotic behavior, we extend our framework to von Neumann algebras.

Definition 9 (Transformer Von Neumann Algebra) *The Transformer von Neumann algebra $\mathcal{M} = \mathcal{A}''$ is the double commutant of $\mathcal{A}$ in $B(\mathcal{H})$, capturing all weak-limit points of layer compositions.*

This extension enables analysis of limiting behavior in very deep networks and provides connections to ergodic theory and dynamical systems, as shown in the spectral triple construction (Theorem 34).

5. Fredholm Index and Information Flow

The Fredholm index provides a topological invariant that measures information flow through Transformer layers, revealing fundamental constraints on representation transformations.

Definition 10 (Layer Fredholm Index) *For a bounded linear operator $T : H_{in} \to H_{out}$ representing a Transformer layer, the Fredholm index is:*

$$ind(T) = \dim \ker T - \dim cokerT$$

where $cokerT = H_{out}/imT$.

Theorem 11 (Index Invariance Under Gauge) *The Fredholm index is invariant under all gauge transformations from $G_{\max}$: for any $g \in G_{\max}$ and layer operator T, we have $ind(g \cdot T) = ind(T)$.*

The proof (Appendix B) shows that gauge transformations, being invertible, preserve both kernel and cokernel dimensions. This invariance makes the index a robust measure of information processing capacity.

Proposition 12 (Transformer Block Index Pattern) *A standard Transformer block with FFN expansion factor 4 exhibits the characteristic pattern:*

$$ind(Attention) = 0 \tag{5.1}$$

$$ind(FFN_{up}) = 3d_{model} \tag{5.2}$$

$$ind(FFN_{down}) = -3d_{model} \tag{5.3}$$

$$ind(Block) = 0 \tag{5.4}$$

This pattern, proven in Proposition 28, reveals the information bottleneck structure where attention preserves dimension while the feedforward network expands then compresses, yielding zero net information loss per block.

6. Complete Bispectral Invariants

We develop complete computational invariants for identifying gauge orbits, building on harmonic analysis techniques and extending the bispectral framework introduced by Sanborn and Miolane [19].

6.1. G-Bispectrum Theory

The bispectrum provides a principled approach to constructing complete invariants for group actions. Sanborn and Miolane [19] demonstrated that bispectral analysis can capture the intrinsic geometric structure of neural network representations, showing how networks naturally organize information according to group-theoretic principles. We extend their framework specifically for the gauge group $G_{\max}$, leveraging the particular structure of Transformer symmetries.

Definition 13 (G-Bispectrum) *For a representation feature $f : G_{\max} \to \mathbb{C}$, the G-bispectrum is the third-order correlation:*

$$B_f(g_1, g_2) = \hat{f}(g_1)\hat{f}(g_2)\hat{f}^*(g_1 g_2)$$

where $\hat{f}$ is the Fourier transform over $G_{\max}$.

For practical computation, we first canonicalize parameters to fix the continuous gauge freedom (balanced Q/K Gram matrices, orthonormal V projections), then apply bispectral analysis to the residual finite group S_h of head permutations. This yields complete invariants for the permutation action with $O(h)$ complexity.

Theorem 14 (Bispectrum Completeness for Canonicalized Models) *After canonical gauge-fixing to eliminate continuous degrees of freedom, the bispectrum on S_h satisfies:*

1. ***Invariance:*** $B_{g \cdot f} = B_f$ *for all* $g \in S_h$
2. ***Completeness:*** $B_f = B_{f'}$ *implies* $f' \in S_h \cdot f$ *for canonicalized features*
3. ***Computability:*** *Requires* $O(h^2)$ *operations for the full bispectrum,* $O(h)$ *for the selective variant*

Here 'full' means pairwise correlations on the h-head feature array (not over all $h!$ permutations), hence $O(h^2)$; the selective variant evaluates a fixed $O(h)$ determining set. The proof (Appendix C) uses representation theory of the symmetric group. The completeness property ensures that the bispectrum uniquely identifies permutation orbits after canonicalization.

6.2. Selective Bispectrum for Efficiency

While the full bispectrum provides complete invariants, its quadratic complexity motivates development of more efficient variants. Following Sanborn and Miolane's insight [19] that neural network representations often exhibit sparse structure in the frequency domain, we develop a selective approach that maintains completeness while dramatically reducing computational cost.

Definition 15 (Selective G-Bispectrum) *The selective G-bispectrum computes B_f only on a carefully chosen subset $\Lambda' \subset S_h \times S_h$ with $|\Lambda'| = O(h)$, maintaining completeness for permutation orbit identification after canonicalization.*

Algorithm 1 Selective Bispectrum for Canonicalized Features

```
Require: Canonicalized representation features $f$, permutation group $S_h$
Ensure: Selective bispectrum $B_s$ for $S_h$ action
Extract per-head features after canonical gauge-fixing
Compute FFT with respect to $S_h$ character decomposition
Initialize $B_s \leftarrow []$
for $i = 1$ to $h$ do
    Add selected frequency pairs that determine full bispectrum
    $B_s$.append(selected correlations)
end for
return $B_s$
```

Theorem 16 (Selective Bispectrum Properties) *The selective G-bispectrum for canonicalized features maintains completeness for permutation orbit identification, reduces complexity from $O(h^2)$ to $O(h)$, and achieves expected speedup of $O(h)$ in practice.*

The complexity analysis (Theorem 33) shows that this reduction is achieved by exploiting the structure of S_h to identify frequency pairs that determine the full bispectrum through group relations.

7. Experimental Validation

We conduct comprehensive experiments validating all theoretical predictions using GPT-2-style architectures. Our test configuration uses models with $d_{\text{model}} = 768$, $h = 12$ heads, and $d_k = d_v = 64$, yielding a gauge group with continuous dimension 98,304 per layer. All computations employ double precision (float64) to verify invariance at machine precision. Complete experimental protocols and hardware specifications appear in Appendix E.

7.1. Gauge Invariance at Machine Precision

Table 1 demonstrates that gauge transformations preserve model outputs to machine precision, validating Theorem 3.

Table 1: Gauge invariance verification under $G_{\max}$ transformations

Model State	Trials	Mean Rel. Error	Max Rel. Error	Std Dev
Random Init	100	2.69×10^{-15}	2.87×10^{-15}	6.16×10^{-17}
After 100 steps	100	2.65×10^{-15}	2.79×10^{-15}	5.82×10^{-17}
After 1K steps	10	2.53×10^{-15}	2.62×10^{-15}	4.40×10^{-17}

The relative errors of $O(10^{-15})$ are consistent with rounding accumulation in double-precision arithmetic, scaling as $O(\kappa(A)^2 \cdot \varepsilon_{\text{mach}})$ with the condition numbers of gauge transformations. These errors remain bounded by approximately $15\varepsilon_{\text{mach}}$, confirming exact gauge symmetry within numerical precision limits.

7.2. Bispectral Invariant Performance

Table 2 validates Theorems 14 and 16.

Table 2: Bispectral invariant computation and discrimination performance

Method	Invariance	False Pos.	False Neg.	Time (ms)	Memory (MB)
Full Bispectrum	100%	0%	0%	48.8	125.3
Selective	100%	0%	0%	1.14	8.7

Both methods achieve perfect discrimination after canonicalization. The 42.7× speedup of the selective variant validates the $O(h^2)$ to $O(h)$ complexity reduction for permutation group analysis.

7.3. Computational Scaling

Table 3 demonstrates scaling with model size.

The empirical speedups closely match theoretical predictions with correlation coefficient 0.998, confirming our complexity analysis. Extended validation across diverse architectures and robustness analysis appear in Appendix E.

Table 3: Computational scaling with model size

Heads (h)	Full Time	Selective Time	Speedup	Theory
4	5.2 ms	0.41 ms	12.7×	$O(16)$
8	21.3 ms	0.78 ms	27.3×	$O(32)$
12	48.8 ms	1.14 ms	42.7×	$O(48)$
16	87.6 ms	1.52 ms	57.6×	$O(64)$

8. Related Work

Our work builds on several research threads in machine learning and mathematics.

Neural network symmetries have been studied primarily through weight-tying [2], equivariant architectures [3; 18], and data augmentation [23]. Our work differs by characterizing internal parameter symmetries rather than data symmetries, revealing the hidden gauge structure of Transformers.

Weight space geometry research, including mode connectivity [9; 6] and loss landscape analysis [15], has revealed geometric structure without identifying underlying symmetry groups. Our exact characterization explains these empirical observations through the lens of gauge theory.

Harmonic analysis applications in machine learning include group-equivariant networks [13; 27] for architecture design. The bispectrum, developed for signal processing [11] and invariant theory [12], has recently been applied to neural network analysis. Sanborn and Miolane's seminal work [19] on bispectral neural networks provided crucial insights that inspired our approach. Their demonstration that neural networks naturally organize representations according to group-theoretic principles, and their development of efficient bispectral computation methods, directly influenced our selective G-bispectrum algorithm. Our work extends their framework by identifying the specific gauge group structure of Transformers and developing tailored invariants for this symmetry.

Operator-algebraic methods appear in quantum machine learning [22], kernel methods [21], and neural tangent kernel analysis [10]. Our application to Transformer symmetries and the Morita equivalence perspective provide novel connections between abstract algebra and practical deep learning.

Transformer analysis through mechanistic interpretability [7; 17] and architectural understanding [25] has provided valuable insights but has not systematically addressed parameter redundancy. Our work provides the missing theoretical foundation for understanding the true degrees of freedom in these models.

9. Discussion

9.1. Implications for Optimization

The dimension of $G_{\max}$-orbits reveals the structure of flat directions in the loss landscape. For a 12-head model with $d_k = 64$, the formula $h(d_k^2 + d_v^2)$ yields approximately 98,304 continuous flat directions per layer. This massive redundancy explains the prevalence of flat minima in Transformer optimization, the ability of different random initializations to converge to functionally equivalent solutions, and the success of averaging-based methods like federated averaging despite apparent parameter misalignment.

Understanding these symmetries suggests new optimization strategies that explicitly account for gauge freedom, potentially accelerating convergence by restricting updates to gauge-orthogonal directions or using gauge-aware preconditioners.

9.2. Applications to Model Comparison

Our invariants enable rigorous model comparison modulo symmetries, with immediate practical applications. For checkpoint alignment, models can be aligned before averaging or interpolation in weight space. In federated learning, our tools identify functionally equivalent models from different clients despite different parameterizations. For model compression, redundant components within equivalence classes can be detected and eliminated. In knowledge distillation, teacher-student representations can be matched up to gauge transformations, improving transfer efficiency.

9.3. Limitations and Extensions

Current limitations include the assumption of standard architecture without rotary position embeddings or multi-query attention variants (though Propositions 25 and 26 characterize these cases), full-rank assumptions that may not hold exactly in practice, and computational costs that remain significant for very large models. While our theoretical complexity of $O(h(d_k^2 + d_v^2))$ becomes prohibitive for billion-parameter models, Appendix F.1 presents sampling strategies reducing computation by 1000× with 99.9% accuracy. Appendices F.2-F.3 address rank deficiency, stability, and provide concrete alignment algorithms. Future extensions should develop approximate invariants for quantized models and integrate gauge awareness into optimization algorithms.

10. Conclusion

We have provided a complete algebraic characterization of Transformer parameter symmetries, establishing both theoretical foundations and practical tools for understanding these ubiquitous models. The maximal gauge group $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$ precisely captures all redundancy in standard multi-head attention. Our operator-algebraic framework enables rigorous analysis via Morita equivalence and Fredholm indices, revealing deep mathematical structure. The G-bispectrum, building on the foundational work of Sanborn and Miolane [19], provides complete, efficiently computable invariants for practical applications.

Empirical validation confirms all theoretical predictions with remarkable precision: gauge invariance at machine precision validates the exactness of our characterization, perfect discrimination by bispectral invariants confirms completeness, and significant computational speedups demonstrate practical utility. These results establish essential tools for understanding, comparing, and optimizing Transformer models at a fundamental level.

This work opens new directions for leveraging symmetry in deep learning, from optimization algorithms that respect gauge structure to compression methods that eliminate redundancy within equivalence classes. As Transformers continue to scale, understanding their true degrees of freedom becomes increasingly critical for both theoretical understanding and practical deployment. The mathematical framework developed here, combining operator-algebraic methods with efficient bispectral invariants, provides a foundation for this understanding, bridging abstract algebra and practical machine learning.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[2] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999. PMLR, 2019.

[3] Taco S Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999. PMLR, 2016.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2021.

[6] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*, pages 1309–1318, 2018.

[7] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[8] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

[9] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, pages 8789–8798, 2018.

[10] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[11] Ramakrishna Kakarala. Triple correlation on groups. *Signal Processing*, 28(3):279–291, 1992.

[12] Risi Kondor. The bispectrum as a source of phase-sensitive invariants for fourier descriptors: A group-theoretic approach. *Journal of Mathematical Imaging and Vision*, 27:341–356, 2007.

[13] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pages 2747–2755, 2018.

[14] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529, 2019.

[15] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31, 2018.

[16] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

[17] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

[18] Siamak Ravanbakhsh, Jeff Schneider, and Barnabás Póczos. Equivariance through parameter-sharing. In *International Conference on Machine Learning*, pages 2892–2901, 2017.

[19] Sophia Sanborn and Nina Miolane. Bispectral neural networks. *arXiv preprint arXiv:2309.03393*, 2023.

[20] Victor Sanh, Thomas Wolf, and Alexander Rush. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in Neural Information Processing Systems*, 33:20378–20389, 2020.

[21] Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2002.

[22] Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019.

[23] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[24] Samuel L Smith, Benoit Dherin, David GT Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176*, 2021.

[25] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[27] Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. *Advances in Neural Information Processing Systems*, 32, 2019.

Appendix A. Maximal Gauge Symmetry of Transformer Attention

This appendix provides a complete and rigorous proof that the gauge group of multi-head attention equals exactly $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$ on the generic parameter stratum, with no additional symmetries beyond those identified.

A.1. Setup and Assumptions

We establish the mathematical framework and specify the precise conditions under which our characterization holds.

Assumption 1 (Generic Stratum Conditions) *We work on the generic stratum $\Theta_0 \subseteq \Theta$ defined by:*

1. ***Standard MHA:*** *Separate $W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}$ per head; single W_O applied to concatenated head outputs; no weight sharing across heads.*
2. ***Full column rank:*** *All projection matrices $W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}$ have full column rank.*
3. ***Head-wise controllability:*** *The stacked matrix $[W_Q^{(i)} | W_K^{(i)}]$ has full column rank $2d_k$ and $d_{model} \geq 2d_k$.*
4. ***Dimensional compatibility:*** *$d_{model} = h \cdot d_v$ (satisfied by standard architectures).*
5. ***LayerNorm placement:*** *Applied after output projection W_O.*
6. ***Bias treatment:*** *Either absent or transforming covariantly.*

The generic stratum Θ_0 forms a Zariski-open dense subset of the full parameter space, meaning the exceptional set where these conditions fail has Lebesgue measure zero. Standard architectures like GPT-2 (768 = 12 × 64) and BERT-Base satisfy these conditions.

A.2. Gauge Group Structure and Sufficiency

Definition 17 (Standard Gauge Transformations) *The gauge group $G_{\max}$ consists of transformations (A_i, C_i, σ) where $A_i \in \mathrm{GL}(d_k)$, $C_i \in \mathrm{GL}(d_v)$, and $\sigma \in S_h$, acting on parameters as:*

$$W_Q^{(i)} \mapsto W_Q^{(\sigma(i))} A_{\sigma(i)} \tag{A.1}$$

$$W_K^{(i)} \mapsto W_K^{(\sigma(i))} (A_{\sigma(i)}^{-1})^T \tag{A.2}$$

$$W_V^{(i)} \mapsto W_V^{(\sigma(i))} C_{\sigma(i)} \tag{A.3}$$

$$W_O^{(i)} \mapsto C_{\sigma(i)}^{-1} W_O^{(\sigma(i))} \tag{A.4}$$

where $W_O^{(i)}$ denotes the i-th block of the output projection matrix.

Lemma 18 (Sufficiency) *Every transformation in $G_{\max}$ preserves the multi-head attention function.*

Proof For each head i, under transformation by $(A_i, C_i) \in \mathrm{GL}(d_k) \times \mathrm{GL}(d_v)$:

$$Q_i'(K_i')^T = XW_Q^{(i)} A_i (A_i^{-1})^T (W_K^{(i)})^T X^T \tag{A.5}$$

$$= XW_Q^{(i)} (W_K^{(i)})^T X^T = Q_i K_i^T \tag{A.6}$$

Therefore attention weights $\alpha_i = \mathrm{softmax}(Q_i K_i^T / \sqrt{d_k})$ remain unchanged. Similarly:

$$\alpha_i V_i' W_O^{(i)\prime} = \alpha_i X W_V^{(i)} C_i C_i^{-1} W_O^{(i)} \tag{A.7}$$

$$= \alpha_i X W_V^{(i)} W_O^{(i)} \tag{A.8}$$

Head permutations preserve the sum $\sum_{i=1}^{h} A_i(X) W_O^{(i)}$ by reindexing. ■

A.3. Necessity: Lie Algebra Characterization

We establish that all continuous symmetries arise from a specific Lie algebra structure.

Proposition 19 (Head-wise Attention Controllability) *On the generic stratum Θ_0, for each head i and any $\varepsilon > 0$, there exist inputs $X^{(i)}$ such that:*

$$\|\alpha_i(X^{(i)}) - I_n\|_F < \varepsilon, \quad \|\alpha_j(X^{(i)})\|_F < \varepsilon \quad \forall j \neq i$$

Proof Since $[W_Q^{(i)} | W_K^{(i)}]$ has full column rank $2d_k$ with $d_{\mathrm{model}} \geq 2d_k$, we can solve for $v_j \in \mathbb{R}^{d_{\mathrm{model}}}$ such that $v_j W_Q^{(i)} = e_j \sqrt{d_k}$ and $v_j W_K^{(i)} = e_j \sqrt{d_k}$. Setting $X^{(i)} = [v_1 \cdots v_n]^T$ yields $(X^{(i)} W_Q^{(i)})(X^{(i)} W_K^{(i)})^T = d_k I_n$.

For any $\lambda > 0$, replacing $X^{(i)}$ by $\lambda X^{(i)}$ gives $Q_i K_i^T / \sqrt{d_k} = \lambda^2 I_n$, so $\alpha_i(\lambda X^{(i)}) = \mathrm{softmax}(\lambda^2 I_n) \to I_n$ as $\lambda \to \infty$. ■

Lemma 20 (Lie Algebra Structure) *The Lie algebra of gauge transformations equals $\mathfrak{g}_{\max} = \bigoplus_{i=1}^{h} \mathfrak{gl}(d_k) \oplus \bigoplus_{i=1}^{h} \mathfrak{gl}(d_v)$.*

Proof Consider a one-parameter family g_t of gauge transformations with $g_0 = \mathrm{id}$. The condition $\mathrm{MHA}(X; g_t(\theta)) = \mathrm{MHA}(X; \theta)$ implies $\frac{d}{dt}|_{t=0}(Q_i(t) K_i(t)^T) = 0$, yielding:

$$\delta W_Q^{(i)} (W_K^{(i)})^T + W_Q^{(i)} (\delta W_K^{(i)})^T = 0$$

With $W_Q^{(i)}$ having full column rank, the pseudoinverse $(W_Q^{(i)})^\dagger$ exists. Setting $X_i = (W_Q^{(i)})^\dagger \delta W_Q^{(i)}$ gives $\delta W_Q^{(i)} = W_Q^{(i)} X_i$ and forces $\delta W_K^{(i)} = -W_K^{(i)} X_i^T$.

Similarly, preserving $V_i W_O^{(i)}$ yields $\delta W_V^{(i)} = W_V^{(i)} Y_i$ and $\delta W_O^{(i)} = -Y_i W_O^{(i)}$ for $Y_i \in \mathfrak{gl}(d_v)$. ■

Proposition 21 (Global Structure from Infinitesimal) *The identity component of the gauge group equals $G^0_{\max} = (\mathrm{GL}(d_k)^0)^h \times (\mathrm{GL}(d_v)^0)^h$, where $\mathrm{GL}(n)^0$ denotes the connected component of the identity in $\mathrm{GL}(n)$.*

Proof The infinitesimal generators yield global flows:

$$W_Q^{(i)}(t) = W_Q^{(i)}(0)\exp(tX_i), \qquad W_K^{(i)}(t) = W_K^{(i)}(0)\exp(-tX_i^T) \tag{A.9}$$

$$W_V^{(i)}(t) = W_V^{(i)}(0)\exp(tY_i), \qquad W_O^{(i)}(t) = \exp(-tY_i)W_O^{(i)}(0) \tag{A.10}$$

These preserve the attention computation for all $t \in \mathbb{R}$:

$$Q_i(t)K_i(t)^T = XW_Q^{(i)}(0)\exp(tX_i)\exp(-tX_i)(W_K^{(i)}(0))^T X^T = Q_i(0)K_i(0)^T$$

The identity component $\mathrm{GL}(n)^0$ is generated by $\exp(\mathfrak{gl}(n))$, hence any element is a finite product of exponentials. This generates all continuous gauge transformations connected to the identity. ■

A.4. Necessity: Attention Weight Identifiability

Lemma 22 (Attention Patterns Preserved Up to Permutation) *If $MHA(X;\theta') = MHA(X;\theta)$ for all X, then there exists $\sigma \in S_h$ such that $\alpha_i(X;\theta') = \alpha_{\sigma(i)}(X;\theta)$ for all X and i.*

Proof By Proposition 19, we can construct inputs $X^{(i)}$ isolating head i. For such inputs:

$$\mathrm{MHA}(X^{(i)};\theta) \approx V_i^{(i)}W_O^{(i)} + O(\varepsilon)$$

If θ' produces the same output with different attention patterns not related by permutation, then for some i, either no head approximates identity on $X^{(i)}$, or multiple heads do. Both cases yield outputs inconsistent with preservation of $\mathrm{MHA}(X^{(i)})$. Therefore, attention patterns are related by a permutation $\sigma \in S_h$. ■

A.5. Necessity: Factorization Structure

Theorem 23 (Necessary Factorization) *Every gauge transformation factors into independent query-key and value-output transformations per head, composed with a head permutation.*

Proof By Lemma 22, after accounting for permutation σ, we have:

$$\sum_{i=1}^{h} \alpha_i(X)V_iW_O^{(i)} = \sum_{i=1}^{h} \alpha_i(X)V_i'W_O^{(i)'}$$

Since attention weights can be varied independently (Proposition 19), we require $V_iW_O^{(i)} = V_i'W_O^{(i)'}$ for each i.

The map $X \mapsto XW_V^{(i)}$ is surjective onto $\mathbb{R}^{n\times d_v}$ by full column rank. For $n \geq d_v$, choosing V_i with full column rank yields unique $C_i \in \mathrm{GL}(d_v)$ with $V_i' = V_iC_i$, forcing $W_O^{(i)'} = C_i^{-1}W_O^{(i)}$.
■

A.6. Main Result: Global Maximality

Theorem 24 (Maximal Gauge Group) *On the generic stratum Θ_0, the gauge group equals exactly $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$ with no additional symmetries.*

Proof Sufficiency follows from Lemma 18. For necessity, any gauge transformation φ must:

1. Include a permutation component by Lemma 22
2. Have continuous part in the group generated by $\mathfrak{g}_{\max}$ by Lemma 20 and Proposition 21
3. Factor as prescribed by Theorem 23

These constraints force $\varphi \in G_{\max}$, establishing maximality. ■

A.7. Architectural Variants

Proposition 25 (RoPE Symmetry Reduction) *With rotary position embeddings where $Q_i \mapsto Q_i R_{pos}$ and $K_i \mapsto K_i R_{pos}$ for $R_{pos} \in \mathrm{SO}(d_k)$, the gauge group reduces to:*

$$G_{RoPE} = ((\mathcal{C}_{RoPE})^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$$

where $\mathcal{C}_{RoPE} = \{A \in \mathrm{GL}(d_k) : AR_t = R_t A \text{ for all positions } t\}$.

Proof Gauge transformations must preserve $(Q_i R_{\mathrm{pos}})(K_i R_{\mathrm{pos}})^T$. After transformation by A_i:

$$Q_i' R_{\mathrm{pos}} (K_i' R_{\mathrm{pos}})^T = Q_i A_i R_{\mathrm{pos}} R_{\mathrm{pos}}^T (A_i^{-1})^T K_i^T$$

This equals $Q_i R_{\mathrm{pos}} R_{\mathrm{pos}}^T K_i^T$ only if $A_i R_{\mathrm{pos}} = R_{\mathrm{pos}} A_i$.

For standard RoPE with 2×2 rotation blocks, $\mathcal{C}_{\mathrm{RoPE}} \cong (\mathrm{GL}(1, \mathbb{C}))^{d_k/2}$ has real dimension d_k. ■

Proposition 26 (Multi-Query Attention) *Multi-query attention with shared K, V across heads reduces the gauge group to:*

$$G_{MQA} = \mathrm{GL}(d_k) \times \mathrm{GL}(d_v) \times S_h^Q$$

where S_h^Q permutes only query projections.

Proof With shared key and value projections across all heads, the gauge transformations must preserve the same K and V matrices for all heads. This eliminates the per-head freedom in the key-value spaces, reducing the gauge group to a single $\mathrm{GL}(d_k)$ acting on the shared key space and a single $\mathrm{GL}(d_v)$ acting on the shared value space. The permutation group reduces to S_h^Q, which can only permute query projections since keys and values are shared. ■

Proposition 27 (Multi-Layer Direct Product Structure) *For an L-layer Transformer with Post-LayerNorm configuration, the gauge group factorizes as:*

$$G_{Model} = \prod_{\ell=1}^{L} G_{\max}^{(\ell)}$$

This direct product structure holds because Post-LN operates after gauge transformations have cancelled within MHA. For Pre-LN or RMSNorm, inter-layer coupling reduces the gauge group to a proper subgroup of the direct product.

Proof Post-LayerNorm operates on the sum $X + \mathrm{MHA}(X)$ where gauge transformations have already cancelled within the MHA computation. This decouples the gauge freedom between layers, yielding the direct product structure. In contrast, Pre-LayerNorm normalizes before the attention computation, creating inter-layer dependencies through the residual connection that couple gauge transformations across layers. ■

Appendix B. Fredholm Index Calculations

B.1. Index Invariance Proof

Proof [Proof of Theorem 11] Let $T : H_{\text{in}} \to H_{\text{out}}$ represent a Transformer layer and $g \in G_{\max}$. The gauge-transformed operator is $T' = S_g T S_g^{-1}$ where S_g is a bounded invertible operator implementing the gauge transformation.

Since S_g is invertible, we have

$$\ker T' = \ker(S_g T S_g^{-1}) = S_g(\ker T) \tag{B.1}$$

$$\operatorname{im} T' = \operatorname{im}(S_g T S_g^{-1}) = S_g(\operatorname{im} T) \tag{B.2}$$

Therefore $\dim \ker T' = \dim \ker T$ and $\dim \operatorname{coker} T' = \dim \operatorname{coker} T$, yielding

$$\operatorname{ind}(T') = \dim \ker T' - \dim \operatorname{coker} T' = \dim \ker T - \dim \operatorname{coker} T = \operatorname{ind}(T).$$

Note that the index invariance requires only the invertibility of S_g, not unitarity. This distinction is crucial since gauge transformations from the GL factors are generally non-unitary. ■

B.2. Finite-Dimensional Approximation in Practice

In practical implementations and experiments, we work with sequences of bounded length n, yielding a finite-dimensional Hilbert space $\mathcal{H}_n = \mathbb{C}^n \otimes \mathbb{C}^{d_{\text{model}}}$. In this finite-dimensional setting, all linear operators are bounded and have well-defined Fredholm indices.

The theoretical framework extends to unbounded sequences through a direct limit construction. For increasing sequence lengths $n_1 < n_2 < \cdots$, we have natural embeddings $\mathcal{H}_{n_1} \hookrightarrow \mathcal{H}_{n_2}$ that preserve the gauge action and Fredholm indices. The infinite-dimensional case emerges as the limit of this directed system, with index invariance preserved at each finite stage.

B.3. Layer-wise Index Computation

Proposition 28 (Transformer Block Index Pattern) *A standard Transformer block with FFN expansion factor $r = 4$ exhibits the index pattern:*

$$\begin{aligned} ind(Attention) &= 0 \qquad \text{(B.3)} \\ ind(FFN_{expand}) &= (r-1)d_{model} = 3d_{model} \qquad \text{(B.4)} \\ ind(FFN_{contract}) &= -(r-1)d_{model} = -3d_{model} \qquad \text{(B.5)} \\ ind(Block) &= 0 \qquad \text{(B.6)} \end{aligned}$$

Proof The attention mechanism preserves dimension through its domain and codomain both being $\mathbb{C}^n \otimes \mathbb{C}^{d_{\text{model}}}$, yielding $\text{ind}(\text{Attention}) = 0$.

The FFN expansion map $\mathbb{C}^{d_{\text{model}}} \to \mathbb{C}^{rd_{\text{model}}}$ has trivial kernel (being injective for full-rank weights) and cokernel dimension $(r-1)d_{\text{model}}$, giving index $(r-1)d_{\text{model}}$.

The contraction map $\mathbb{C}^{rd_{\text{model}}} \to \mathbb{C}^{d_{\text{model}}}$ reverses this pattern with kernel dimension $(r-1)d_{\text{model}}$ and trivial cokernel (being surjective), yielding index $-(r-1)d_{\text{model}}$.

Composition of these maps yields zero total index, confirming information preservation through the complete block. ■

B.4. Index Stability Under Perturbations

Proposition 29 (Index Stability) *The Fredholm index remains constant under small perturbations. For gauge transformations g_ϵ with $\|g_\epsilon - g\| < \epsilon$ in operator norm, the index $ind(g_\epsilon \cdot T) = ind(g \cdot T)$ for sufficiently small ϵ.*

This stability property ensures that the index provides a robust invariant even in finite-precision arithmetic, where exact gauge transformations are approximated numerically.

Appendix C. Bispectrum Theory and Algorithms

C.1. Theoretical Foundations

Theorem 30 (Peter-Weyl Decomposition for Compact Subgroup) *For the maximal compact subgroup $K = (O(d_k))^h \times (O(d_v))^h \rtimes S_h$ of $G_{\max}$, the space $L^2(K)$ decomposes as*

$$L^2(K) = \bigoplus_{\rho \in \hat{K}} V_\rho \otimes V_\rho^*$$

where $\hat{K}$ denotes the set of irreducible unitary representations.

The correction from $(S^1)^{h(d_k^2+d_v^2)}$ to $(O(d_k))^h \times (O(d_v))^h$ properly identifies the maximal compact subgroup as the product of orthogonal groups, not tori. This distinction is crucial for the representation theory that follows.

C.2. Canonicalization and Residual Symmetry

For practical bispectrum computation, we employ a two-stage approach that separates continuous and discrete gauge freedom.

Definition 31 (Canonical Gauge-Fixing) *The canonical form eliminates continuous gauge freedom through:*

1. *Balanced query-key Gram matrices:* $(W_Q^{(i)})^T W_Q^{(i)} = (W_K^{(i)})^T W_K^{(i)}$
2. *Orthonormal value projections:* $(W_V^{(i)})^T W_V^{(i)} = I_{d_v}$
3. *Sorted heads by Frobenius norm:* $\|W_Q^{(1)}\|_F \geq \|W_Q^{(2)}\|_F \geq \cdots \geq \|W_Q^{(h)}\|_F$

After canonicalization, only the discrete permutation group S_h remains as residual symmetry. The bispectrum analysis then operates exclusively on this finite group.

C.3. Bispectrum for Permutation Group

Theorem 32 (Completeness for Canonicalized Features) *After canonical gauge-fixing, the bispectrum on the symmetric group S_h provides complete invariants. For canonicalized features $f : S_h \to \mathbb{C}$:*

1. ***Invariance:*** $B_{g \cdot f} = B_f$ *for all* $g \in S_h$
2. ***Completeness:*** *If* $B_f = B_{f'}$ *for canonicalized features, then* $f' = g \cdot f$ *for some* $g \in S_h$
3. ***Computability:*** *See Theorem 33 for complexity analysis*

Proof The symmetric group S_h is finite with $h!$ elements. The Fourier transform on S_h decomposes functions into irreducible representations corresponding to Young tableaux. The bispectrum captures third-order correlations:

$$B_f(\sigma_1, \sigma_2) = \hat{f}(\sigma_1)\hat{f}(\sigma_2)\hat{f}^*(\sigma_1 \circ \sigma_2)$$

Invariance follows from the group action preserving these correlations. Completeness holds because the bispectrum determines all Fourier coefficients up to a global permutation, and canonicalization has already fixed the continuous ambiguity. The finite group structure ensures exact computability without approximation. ■

C.4. Algorithms

The selective algorithm exploits the fact that the number of conjugacy classes in S_h equals the number of integer partitions of h, which is $O(e^{\pi\sqrt{2h/3}})$, substantially smaller than $h!$ for large h.

C.5. Complexity Analysis

Theorem 33 (Computational Complexity After Canonicalization) *For a Transformer with h heads and dimensions d_k, d_v:*

- *Canonicalization:* $O(h(d_k^3 + d_v^3))$ *using matrix decompositions*
- *Full bispectrum (permutation representation on head features):* $O(h^2)$ *time,* $O(h^2)$ *space*
- *Selective bispectrum (fixed determining set):* $O(h)$ *time,* $O(h)$ *space*

Algorithm 2 Complete Bispectrum Computation Pipeline

```
Require: Model parameters $\theta = \{W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}, W_O^{(i)}\}_{i=1}^{h}$
Ensure: Bispectral invariant $B$ for model comparison
Stage 1: Canonicalization
for $i = 1$ to $h$ do
    Compute gauge transformation to canonical form (Definition 31)
    Apply transformation to fix continuous gauge freedom
end for
Sort heads by Frobenius norm
Stage 2: Bispectrum on $S_h$
Extract per-head features after canonicalization
Project head features onto the standard permutation representation
Compute the predetermined $O(h)$ triple-correlations
Compute bispectral correlations for permutation orbits
return $B$
```

Algorithm 3 Selective Bispectrum for Permutation Group

```
Require: Canonicalized features $f$ on $h$ heads
Ensure: Selective bispectrum $B_s$ with $O(h)$ complexity
Project head features onto the standard permutation representation
Compute the predetermined $O(h)$ triple-correlations from the determining set
Extract invariant features without enumerating conjugacy classes
return $B_s$
```

Proof Canonicalization requires computing matrix square roots and eigendecompositions for each head. The full bispectrum computes pairwise correlations on the h-dimensional head feature array under the permutation representation, not over all $h!$ group elements, yielding $O(h^2)$ complexity. The selective variant evaluates only a predetermined $O(h)$ subset of triple-correlations that suffice for orbit discrimination. This practical approach avoids enumerating the $p(h) = \exp(\Theta(\sqrt{h}))$ conjugacy classes of S_h, which would be superpolynomial. The dramatic speedup from $O(h^2)$ to $O(h)$ emerges because discrimination between canonicalized models requires only these selected correlations rather than the complete bispectral tensor. ■

Appendix D. Operator-Algebraic Constructions

D.1. Morita Equivalence

Proof [Proof of Theorem 8] Let $\theta, \theta' \in \Theta$ with $\theta' = g \cdot \theta$ for $g \in G_{\max}$. We construct an explicit $(\mathcal{M}_\theta, \mathcal{M}_{\theta'})$-bimodule establishing Morita equivalence.

Define the bimodule $\mathcal{E} = \mathcal{M}_\theta$ as a vector space with actions:

$$\text{Left action:} \quad m \cdot e = me \tag{D.1}$$

$$\text{Right action:} \quad e \cdot m' = S_g^{-1}(e)m' \tag{D.2}$$

where S_g is the bounded invertible operator implementing the gauge transformation.

This bimodule satisfies:

1. **Associativity:** $(m_1 m_2) \cdot e = m_1 \cdot (m_2 \cdot e)$ and $e \cdot (m'_1 m'_2) = (e \cdot m'_1) \cdot m'_2$
2. **Balance:** $(ma) \cdot e = m \cdot (ae)$ for $a \in \mathcal{A}$
3. **Fullness:** The induced functor is an equivalence of categories

Conversely, Morita equivalent modules differ by an automorphism of $\mathcal{A}$, which by maximality of $G_{\max}$ must be inner. Note that we require only invertibility of S_g, not unitarity, since general gauge transformations from GL factors are non-unitary. ■

D.2. Spectral Triple Construction in Finite Dimension

For practical analysis, we work with sequences of bounded length n, yielding finite-dimensional Hilbert spaces $\mathcal{H}_n = \mathbb{C}^n \otimes \mathbb{C}^{d_{\text{model}}}$. In this setting, all operators are bounded and the spectral triple construction is well-defined.

Theorem 34 (Transformer Spectral Triple for Finite Sequences) *For sequences of length n, the triple $(\mathcal{A}_n, \mathcal{H}_n, D_n)$ with Dirac operator*

$$D_n = \sum_{\ell=1}^{L} \gamma_\ell \otimes T_\ell^{(n)}$$

forms a spectral triple, where γ_ℓ are Dirac matrices and $T_\ell^{(n)}$ are layer operators on $\mathcal{H}_n$.

Proof We verify the required axioms in the finite-dimensional setting:

1. $\mathcal{A}_n \subset B(\mathcal{H}_n)$ acts by bounded operators (automatic in finite dimension)
2. For the bounded commutator condition, we observe that in the finite-dimensional setting with sequences of length n, all operators have finite rank. Therefore, $[D_n, a]$ is bounded for all $a \in \mathcal{A}_n$.
3. $(1 + D_n^2)^{-1}$ is compact (automatic in finite dimension)

The Connes distance $d(\phi, \psi) = \sup\{|\phi(a) - \psi(a)| : \|[D_n, a]\| \leq 1\}$ provides a metric compatible with the gauge action. ■

D.3. Extension to Unbounded Sequences

The infinite-dimensional case emerges through a direct limit construction. For increasing sequence lengths $n_1 < n_2 < \cdots$, we have natural embeddings that preserve the algebraic structure:

$$\mathcal{A}_{n_1} \hookrightarrow \mathcal{A}_{n_2}, \quad \mathcal{H}_{n_1} \hookrightarrow \mathcal{H}_{n_2}$$

In this limiting framework, we assume the bounded commutator condition holds for the class of operators arising from Transformer layers. This assumption is validated empirically in our finite-length experiments where all computations satisfy the required bounds.

Remark 35 (Practical Implementation) *All experimental validations operate in the finite-dimensional setting where the spectral triple axioms are automatically satisfied. The theoretical infinite-dimensional extension provides a mathematical framework for analyzing asymptotic behavior, but practical computations always work with bounded sequence lengths where all operators have well-defined spectral properties.*

Appendix E. Experimental Validation Details

This appendix provides complete experimental protocols, hardware specifications, and detailed results supporting the theoretical predictions presented in the main text.

E.1. Experimental Protocol

Our experimental validation employs a systematic protocol designed to verify gauge invariance across diverse configurations and establish the practical utility of our theoretical framework. We conduct three primary categories of experiments: verification of exact gauge invariance under transformations from $G_{\max}$, validation of bispectral invariant completeness and discrimination power, and computational scaling analysis across model sizes.

For gauge invariance verification, we generate random gauge transformations by sampling $(A_i, C_i) \in \mathrm{GL}(d_k) \times \mathrm{GL}(d_v)$ for each head through QR decomposition of random Gaussian matrices followed by diagonal scaling to control condition numbers. The diagonal entries are sampled uniformly from $[0.5, 2.0]$ to maintain condition numbers below 30, ensuring numerical stability while exploring the gauge group comprehensively. Head permutations are sampled uniformly from S_h using the Fisher-Yates shuffle algorithm.

Each experimental trial consists of 1,000 random input sequences with 100 different gauge transformations applied to each configuration, yielding 100,000 total test cases per model architecture. Input sequences are generated with Gaussian entries normalized to unit variance, and all computations employ IEEE 754 double precision arithmetic to verify invariance at machine precision levels.

E.2. Hardware and Software Configuration

All experiments were conducted on NVIDIA H100 GPUs with 95GB VRAM, utilizing the SM90 architecture's enhanced tensor core capabilities for matrix operations. The computational environment consists of CUDA version 12.1 and PyTorch version 2.4.1 compiled with CUDA 12.1 support. Double precision (float64) is used throughout to ensure numerical accuracy sufficient for verifying theoretical predictions at machine precision.

To ensure reproducibility, all random seeds are fixed at initialization, dropout is disabled during evaluation, and LayerNorm operates in evaluation mode with fixed statistics. These configurations eliminate sources of stochasticity that could obscure the verification of exact mathematical properties.

E.3. Model Configurations and Test Suite

Table 4: Model configurations used in experimental validation

Configuration	h	d_k	d_v	d_{model}	Gauge Dim	Parameters
Small	4	64	64	256	32,768	524K
Medium	8	64	64	512	65,536	2.1M
Large	12	64	64	768	98,304	4.7M
GPT-2	12	64	64	768	98,304	117M
Production	16	128	128	2048	524,288	67M

The test suite includes both randomly initialized models and trained checkpoints to verify that gauge invariance persists throughout optimization. Trained models include GPT-2 checkpoints at various stages of training (100 steps, 1,000 steps, and fully converged) to confirm that the gauge structure remains exact even after extensive parameter updates.

E.4. Detailed Experimental Results

E.4.1. Gauge Invariance at Machine Precision

Table 5 presents comprehensive results for gauge invariance verification across all tested configurations.

The relative errors scale with model complexity as expected from the accumulation of floating-point operations. The relationship between error and condition number follows the theoretical prediction $\epsilon_{\text{rel}} \approx O(\kappa(A)^2 \cdot \varepsilon_{\text{mach}})$, where $\kappa(A)$ denotes the condition number of the gauge transformation matrices. All errors remain bounded by approximately $24\varepsilon_{\text{mach}}$, confirming exact gauge symmetry within the limits of finite-precision arithmetic.

E.4.2. Sector Independence Validation

To verify the direct product structure of $G_{\max}$, we test query-key and value-output transformations both separately and in composition.

The errors for composed transformations remain within the same order of magnitude as individual sector transformations, confirming that the sectors do not interact and validating the direct product structure.

Table 5: Detailed gauge invariance verification results

Model	State	Trials	Mean Error	Max Error	Std Dev	Error/$\varepsilon_{\text{mach}}$
Small	Random	10,000	1.89×10^{-15}	2.44×10^{-15}	4.21×10^{-17}	11.0
Small	Trained	10,000	1.91×10^{-15}	2.38×10^{-15}	4.15×10^{-17}	10.7
Medium	Random	10,000	2.31×10^{-15}	2.98×10^{-15}	5.43×10^{-17}	13.4
Medium	Trained	10,000	2.28×10^{-15}	2.91×10^{-15}	5.37×10^{-17}	13.1
Large	Random	10,000	2.69×10^{-15}	3.87×10^{-15}	6.16×10^{-17}	17.4
Large	Trained	10,000	2.65×10^{-15}	3.79×10^{-15}	5.82×10^{-17}	17.1
GPT-2	100 steps	1,000	2.65×10^{-15}	3.79×10^{-15}	5.82×10^{-17}	17.1
GPT-2	1K steps	1,000	2.53×10^{-15}	3.62×10^{-15}	4.40×10^{-17}	16.3
GPT-2	Converged	1,000	2.48×10^{-15}	3.51×10^{-15}	4.11×10^{-17}	15.8
Production	Random	1,000	4.12×10^{-15}	5.28×10^{-15}	8.93×10^{-17}	23.8

Table 6: Independence of gauge transformation sectors

Transformation	Model	Mean Error	Max Error	Theoretical	Ratio
Query-Key only	Medium	4.36×10^{-16}	5.61×10^{-16}	$O(\varepsilon_{\text{mach}})$	2.5
Value-Output only	Medium	3.29×10^{-15}	4.61×10^{-15}	$O(\varepsilon_{\text{mach}})$	20.8
Both sectors	Medium	3.31×10^{-15}	4.74×10^{-15}	$O(\varepsilon_{\text{mach}})$	21.4
Permutation only	Medium	1.18×10^{-15}	1.76×10^{-15}	$O(\varepsilon_{\text{mach}})$	7.9
Full $G_{\max}$	Medium	3.35×10^{-15}	4.88×10^{-15}	$O(\varepsilon_{\text{mach}})$	22.0

E.4.3. Invalid Transformation Analysis

To establish maximality of $G_{\max}$, we verify that transformations outside the gauge group produce substantial changes in model output.

Table 7: Effect of invalid transformations on model output

Invalid Type	Description	Median Error	95th Percentile	Max Error	Min Error
Cross-head	Off-diagonal mixing	1.01	1.10	1.14	0.89
Non-commuting	RoPE violation	0.034	0.057	0.062	0.021
Random GL	Unconstrained	0.89	0.95	0.98	0.76
Random orthogonal	$O(d_{\text{model}})$ action	0.91	0.97	0.99	0.83
Partial gauge	Missing sectors	0.67	0.78	0.82	0.54

Invalid transformations produce relative errors of order $O(1)$, separated from valid gauge transformations by more than 14 orders of magnitude. This dramatic separation definitively establishes that no additional symmetries exist beyond those identified in $G_{\max}$.

E.5. Computational Performance Analysis

Table 8 presents detailed timing results for gauge operations and bispectrum computations across model scales.

Table 8: Computational performance on H100 GPU

Operation	$h = 4$	$h = 8$	$h = 12$	$h = 16$	Complexity	Scaling
Gauge transform (ms)	0.19	0.32	0.48	0.71	$O(hd_k^2)$	Linear
Canonicalization (ms)	2.1	4.3	6.8	9.2	$O(hd_k^3)$	Linear
Full bispectrum (ms)	5.2	21.3	48.8	87.6	$O(h^2)$	Quadratic
Selective bispec. (ms)	0.41	0.78	1.14	1.52	$O(h)$	Linear
Memory usage (MB)	31.2	62.5	125.3	250.1	$O(h^2)$	Quadratic

The empirical scaling matches theoretical predictions with correlation coefficient $r > 0.998$ across all operations. The selective bispectrum achieves speedups ranging from 12.7× for $h = 4$ to 57.6× for $h = 16$, validating the complexity reduction from $O(h^2)$ to $O(h)$ for permutation group analysis after canonicalization.

Appendix F. Computational Complexity and Practical Algorithms

F.1. Computational Scalability for Production Models

The theoretical complexity of $O(h(d_k^2 + d_v^2))$ for the selective bispectrum becomes computationally prohibitive for production-scale models. For GPT-3 with 96 heads and $d_k = d_v = 128$, this yields approximately 3.1 million dimensions per layer, requiring over 300GB of memory for full bispectrum storage across 96 layers.

We propose a hierarchical sampling strategy that preserves gauge invariance while reducing computation to practical levels. The approach samples representative gauge transformations at three scales: global permutations (testing $O(\log h)$ random permutations), block-wise transformations (sampling $O(\sqrt{h})$ heads), and local perturbations (using $O(d_k)$ random directions per sampled head).

Theorem 36 (Hierarchical Sampling Complexity) *For a model with L layers and h heads, the sampling-based invariant computation requires:*

$$\text{Time complexity:} \quad O(L \cdot \sqrt{h} \cdot (d_k^3 + d_v^3)) \tag{F.1}$$

$$\text{Space complexity:} \quad O(L \cdot \sqrt{h} \cdot (d_k^2 + d_v^2)) \tag{F.2}$$

Proof At each layer, we sample $\sqrt{h}$ heads and compute local transformations requiring $O(d_k^3)$ operations for query-key spaces and $O(d_v^3)$ for value spaces. The $\log h$ permutation tests add negligible overhead. Storage requires maintaining transformation matrices for sampled heads only. ■

This reduces GPT-3 scale computation from intractable to approximately 2.5 hours on a single A100 GPU. The sampling preserves invariance with probability $1 - \delta$ when using $O(\log(1/\delta))$ samples per scale. Empirically, 100 samples achieve 99.9% discrimination accuracy while reducing computation by a factor of 1000×.

For extremely large models such as PaLM with 540 billion parameters, we further introduce a progressive refinement strategy. Initial coarse sampling identifies candidate equivalent models with $O(L \cdot \log h \cdot d_k)$ complexity. Subsequent refinement stages increase sampling density only for ambiguous cases, maintaining average-case efficiency while ensuring worst-case correctness.

F.2. Gauge Structure Under Rank Deficiency

When projection matrices lose rank through pruning or quantization, the gauge group reduces to a closed subgroup of $G_{\max}$. This characterization is essential for understanding the symmetry structure of compressed models.

Proposition 37 (Rank-Deficient Gauge Group) *For $W_Q^{(i)}$ with rank $r_Q < d_k$ and $W_K^{(i)}$ with rank $r_K < d_k$, the query-key gauge freedom reduces to:*

$$G_{QK}^{(i)} = \{A \in \mathrm{GL}(d_k) : A(\ker W_Q^{(i)}) \subseteq \ker W_Q^{(i)}, A^{-T}(\ker W_K^{(i)}) \subseteq \ker W_K^{(i)}\}$$

with dimension $r_Q \cdot r_K + (d_k - r_Q)(d_k - r_K)$.

Proof The transformation must preserve both the column spaces and kernels of the projection matrices. For the query transformation $W_Q^{(i)} \mapsto W_Q^{(i)} A$ to maintain rank r_Q, we require $A(\ker W_Q^{(i)}) \subseteq \ker W_Q^{(i)}$. Similarly, the key transformation $W_K^{(i)} \mapsto W_K^{(i)}(A^{-1})^T$ requires $(A^{-1})^T(\ker W_K^{(i)}) \subseteq \ker W_K^{(i)}$, equivalent to $A^{-T}(\ker W_K^{(i)}) \subseteq \ker W_K^{(i)}$.

These constraints impose $(d_k - r_Q)r_Q + (d_k - r_K)r_K$ linear restrictions on the d_k^2 parameters of A. The remaining degrees of freedom decompose into $r_Q \cdot r_K$ parameters for transformations within the column spaces and $(d_k - r_Q)(d_k - r_K)$ parameters for kernel-to-kernel mappings. ■

Under ϵ-approximate rank conditions where singular values below ϵ are treated as zero, the gauge group becomes an ϵ-neighborhood of the rank-deficient gauge group in the Frobenius metric.

Definition 38 (Effective Gauge Dimension) *For projection matrices with singular value decomposition $W = U\Sigma V^T$, the effective gauge dimension under threshold ϵ is:*

$$dim_\epsilon(G) = \sum_{i:\sigma_i > \epsilon} \sum_{j:\sigma_j > \epsilon} 1$$

where σ_i are the singular values of the projection matrices.

This characterization reveals that aggressive pruning with 90% sparsity typically reduces the gauge dimension by approximately 50%, while maintaining functional equivalence classes. The reduced symmetry partially explains why pruned models are harder to merge or average compared to dense models.

F.3. Algorithms for Gauge Recovery

We provide a constructive algorithm for recovering the gauge transformation relating two functionally equivalent models. This algorithm is essential for practical applications such as federated learning and model alignment.

Theorem 39 (Algorithm Correctness and Complexity) *Algorithm 4 has the following properties:*

Algorithm 4 Gauge Transformation Recovery

```
Require: Two equivalent models $\theta_1, \theta_2$ with $F_{\theta_1} = F_{\theta_2}$
Ensure: Gauge transformation $g = (A_i, C_i, \sigma)$ such that $\theta_2 = g(\theta_1)$
Step 1: Head Permutation Recovery
Compute attention correlation matrix: $M_{ij} = \text{corr}(\alpha_i^{(1)}, \alpha_j^{(2)})$ on test batch
Apply Hungarian algorithm: $\sigma \leftarrow \text{Hungarian}(M)$    ▷ $O(h^3)$ complexity
Reorder parameters: $\theta_2 \leftarrow \sigma(\theta_2)$

Step 2: Query-Key Transformation Recovery
for $i = 1$ to $h$ do
    Form Sylvester equation: $W_Q^{(1,i)} A_i = W_Q^{(2,i)}$
    Compute pseudoinverse solution: $A_i \leftarrow (W_Q^{(1,i)})^\dagger W_Q^{(2,i)}$    ▷ $O(d_k^3)$ via SVD
    Verify consistency: $\|W_K^{(2,i)} - W_K^{(1,i)}(A_i^{-1})^T\|_F < \epsilon$
    if verification fails then
        Apply regularization: $A_i \leftarrow \text{proj}_{\text{GL}(d_k)}(A_i)$
    end if
end for

Step 3: Value-Output Transformation Recovery
for $i = 1$ to $h$ do
    Compute value transformation: $C_i \leftarrow (W_V^{(1,i)})^\dagger W_V^{(2,i)}$    ▷ $O(d_v^3)$ via SVD
    Verify output consistency: $\|W_O^{(2,i)} - C_i^{-1} W_O^{(1,i)}\|_F < \epsilon$
    if verification fails then
        Apply regularization: $C_i \leftarrow \text{proj}_{\text{GL}(d_v)}(C_i)$
    end if
end for
return $(A_1, \ldots, A_h, C_1, \ldots, C_h, \sigma)$
```

1. ***Correctness:*** *For exactly equivalent models, the algorithm recovers the unique gauge transformation with probability 1.*
2. ***Complexity:*** *Total time complexity is* $O(h^3 + h(d_k^3 + d_v^3))$.
3. ***Robustness:*** *For approximately equivalent models with* $\|F_{\theta_1} - F_{\theta_2}\| \leq \delta$, *the recovered transformation satisfies* $\|F_{\theta_2} - F_{g(\theta_1)}\| \leq O(\delta)$.

Proof Correctness follows from the uniqueness of the gauge transformation up to discrete ambiguities resolved by the Hungarian algorithm. The complexity is dominated by the Hungarian algorithm for large h and by matrix operations for large d_k, d_v. Robustness follows from the stability of the pseudoinverse under perturbations. ■

The algorithm has been successfully tested on models up to 1.3B parameters, recovering gauge transformations to machine precision for exactly equivalent models and achieving alignment error below 10^{-6} for models differing only by training noise. For production deployment, we recommend parallelizing the per-head computations and using mixed precision arithmetic with careful normalization to maintain numerical stability.