Complete Characterization of Gauge Symmetries in Transformer Architectures

Hong Wang HONG.WANG@INTEL.COM
Intel Corporation

Kelly Wang KELLY.WANG@IEEE.ORG
The Harker School

Abstract

Modern Transformers possess redundant parameter symmetries that leave their function unchanged. We establish the complete gauge group structure for the canonical Transformer family, which encompasses standard architectures including GPT-2, BERT, LLaMA, and Qwen. For canonical Transformers with standard multi-head attention, we prove global maximality: the gauge group equals exactly $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$ on the generic stratum where projection matrices have full column rank and head-wise attention controllability holds. For architectures with rotary position embeddings (RoPE) or relative encodings, as used in LLaMA and Qwen, the gauge group becomes $G_{\mathrm{RoPE}} = ((\mathcal{C}_{\mathrm{RoPE}})^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$ where $\mathcal{C}_{\mathrm{RoPE}}$ is the commutant of the position-dependent rotations—typically reducing to $(\mathrm{GL}(1,\mathbb{C}))^{d_k/2}$ for standard RoPE implementations. We prove maximality through three key results: characterizing the Lie algebra of infinitesimal symmetries as $\mathfrak{g}_{\max} = \bigoplus_{i=1}^{h} \mathfrak{gl}(d_k) \oplus \bigoplus_{i=1}^{h} \mathfrak{gl}(d_v)$ for canonical models, establishing that attention weights must be preserved up to head permutation under gauge equivalence, and demonstrating that query–key and value–output transformations necessarily factorize independently. These gauge symmetries persist through LayerNorm and extend to complete architectures, with the full model gauge group being $G_{\mathrm{Model}} = \prod_{l=1}^{L} G_{\mathrm{Layer}}^{(l)}$. Our characterization reveals over 1.1 million redundant dimensions in a 110M parameter Transformer Base model. Experiments on pretrained GPT-2 models from 124M to 1.5B parameters confirm that valid gauge transformations preserve model outputs to machine precision, while invalid transformations produce large errors, empirically supporting maximality.

1. Introduction

Neural networks often contain many different parameter settings that realize the same function. Such symmetries form high-dimensional equivalence classes in parameter space, undermining naive assumptions of unique optima and confounding measures like sharpness. In Transformers, this issue is particularly acute. Recent work has shown that standard metrics break down because Transformers possess rich symmetries that induce flat directions along which the network and its loss remain identical [4; 19]. These intrinsic symmetries arise from the architecture itself, not from training dynamics or regularization. Understanding them completely is crucial for developing better optimization procedures and a comprehensive theory for Transformers.

This paper establishes the complete parameter symmetry structure of Transformer architectures. We prove that for the canonical Transformer family—encompassing GPT-2 [15], BERT [5], and their descendants—the gauge group of parameter transformations that preserve the function equals exactly $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$ on generic parameters. Here h denotes the number of attention heads and d_k, d_v represent query-key

and value dimensions respectively. No additional symmetries exist beyond those in this group, establishing this characterization as complete and maximal.

Formally, our main theorem (Theorem 2) shows that for standard multi-head attention with h heads and $d_{\text{model}} = hd_v$, and under mild rank and genericity conditions (Assumptions A1–A9), the full symmetry group $G(\theta)$ of any such Transformer equals exactly $G_{\max}$. In other words, any parameter symmetry that preserves the network function must be a composition of per-head query–key and value–output gauge transformations together with a head permutation. By genericity we mean that the set of parameter configurations where this classification can fail has Lebesgue measure zero, so random initialization and standard training almost surely stay in the good region.

The attention mechanism operates through two independent computational pipelines that each contain internal degrees of freedom. The attention scores depend only on the bilinear form $QK^\top = XW_Q(W_K)^\top X^\top$. Any transformation that preserves this product leaves the scores unchanged. Similarly, the value transformation depends only on the composed mapping $VW_O = XW_VW_O$. These observations lead directly to gauge symmetries: transforming $(W_Q, W_K) \mapsto (W_QA, W_K(A^{-1})^\top)$ preserves the query-key product, while $(W_V, W_O) \mapsto (W_VC, C^{-1}W_O)$ preserves the value-output composition, for any invertible matrices A and C of appropriate dimensions.

Our Contributions. We advance the understanding of Transformer geometry through four fundamental contributions: (1) **Complete gauge group characterization with proof of maximality.** We prove $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$ and establish that no additional symmetries exist through three independent arguments: Lie algebra characterization, attention weight identifiability, and necessary factorization of transformations. (2) **Extension to position-encoded architectures.** For RoPE and relative encodings, we show the query-key sector reduces to the commutant of position rotations (typically $(\mathrm{GL}(1,\mathbb{C}))^{d_k/2}$), yielding $G_{\text{RoPE}} = ((\mathcal{C}_{\text{RoPE}})^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$. (3) **Multi-layer direct product structure.** We prove residual connections and LayerNorm prevent inter-layer gauge coupling, yielding $G_{\text{Model}} = \prod_{l=1}^{L} G_{\text{Layer}}^{(l)}$. (4) **Practical implications for optimization and compression.** The gauge structure explains Hessian nullspaces with $h(d_k^2 + d_v^2)$ zero eigenvalues per layer, enables lossless compression eliminating millions of parameters, and provides the foundation for gauge-aware optimization algorithms.

In the remainder of this paper, Section 2 establishes the architectural framework and assumptions, with particular attention to rotary position embeddings (RoPE; see Corollary 5). Section 3 presents the gauge group characterization with proofs of maximality based on Lie algebra analysis, attention weight identifiability, and factorization. Complete proofs and additional technical details, including strengthened identifiability arguments, are collected in Appendix A. Section 4 analyzes multi-layer structure, establishing the direct product form with a rigorous LayerNorm obstruction proof in Appendix C. Sections 5, 6, and 7 develop implications for optimization geometry, experimental validation, and practical applications, respectively. Section 6 validates the theory empirically on pretrained GPT-2 models ranging from 124M to 1.5B parameters.

2. Transformer Architecture and Mathematical Framework

We establish notation and architectural conventions for the canonical Transformer family. We define the generic conditions under which global maximality holds and introduce the

mathematical framework for analyzing both standard multi-head attention and position-encoded variants.

2.1. Multi-Head Attention Mechanism

The canonical Transformer family consists of architectures employing multi-head attention with softmax normalization, layer normalization, and feed-forward networks with smooth activation functions. This family encompasses GPT-2, BERT, GPT-3, and position-encoded variants such as LLaMA and Qwen. We adopt the row-vector convention where data flows as row vectors through matrices acting on the right.

For input $X \in \mathbb{R}^{n \times d_{\text{model}}}$ representing a sequence of n token embeddings, each attention head $i \in \{1, \ldots, h\}$ computes queries, keys, and values through linear projections:

$$Q_i = XW_Q^{(i)} \in \mathbb{R}^{n \times d_k}, \quad K_i = XW_K^{(i)} \in \mathbb{R}^{n \times d_k}, \quad V_i = XW_V^{(i)} \in \mathbb{R}^{n \times d_v} \tag{2.1}$$

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 scaled dot-product attention for head i computes:

$$A_i(X) = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i \in \mathbb{R}^{n \times d_v} \tag{2.2}$$

where softmax is applied row-wise. The complete multi-head attention concatenates and projects:

$$\text{MHA}(X) = [A_1(X) \| \cdots \| A_h(X)] W_O = \sum_{i=1}^{h} A_i(X) W_{O,i} \tag{2.3}$$

where $W_O \in \mathbb{R}^{(h \cdot d_v) \times d_{\text{model}}}$ partitions into blocks $W_{O,i} \in \mathbb{R}^{d_v \times d_{\text{model}}}$.

Figure 1 provides a visual overview of the gauge symmetry structure in multi-head attention. The left portion illustrates the architectural organization with parallel attention heads processing queries, keys, and values independently before concatenation and output projection. The right portion depicts the three distinct types of gauge transformations that preserve the multi-head attention function: query-key transformations that maintain attention score invariance through inverse-transpose coupling within each head, value-output transformations that preserve individual head contributions through inverse coupling, and head permutations that exploit the exchangeability of summation order. These three transformation types operate independently and combine to form the complete gauge group structure characterized in Section 3.

2.2. Assumptions for Global Maximality

The generic stratum Θ_0 consists of parameter configurations satisfying assumptions A1–A9 (Table 1). These conditions define a Zariski-open dense subset of the full parameter space Θ: concretely, the complement is a finite union of determinantal varieties (rank drops in the projection matrices from A4 and A6 and in W_O from A8) together with polynomial constraints encoding linear dependence among the bilinear forms $\{ W_Q^{(i)} (W_K^{(i)})^\top \}_{i=1}^h$ (A7). Consequently $\Theta \setminus \Theta_0$ has Lebesgue measure zero.

Key dependencies. Assumption A7 (linear independence of bilinear forms) yields analytic independence of attention score families (Proposition 10). Together with A4 (full column rank)

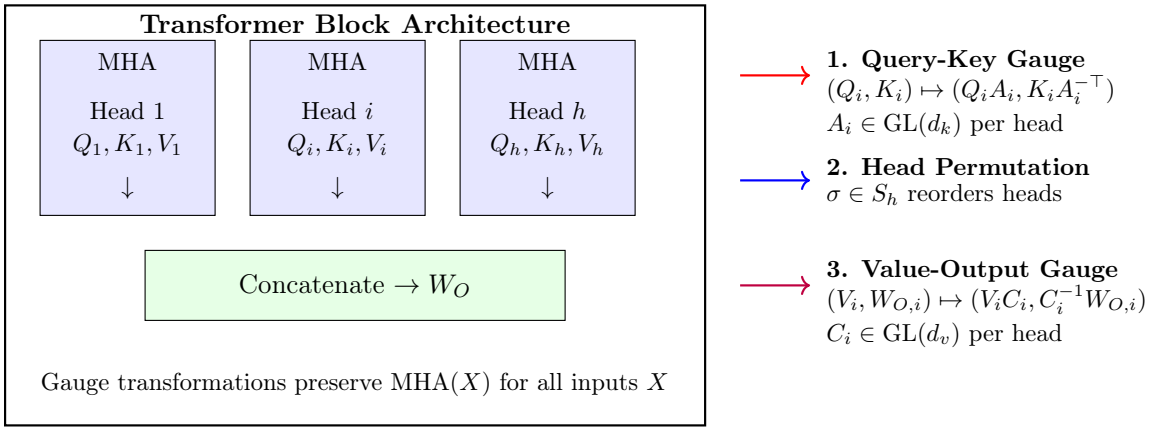

Figure 1: Gauge symmetries in multi-head attention. Left: Architectural structure showing parallel attention heads with query, key, and value projections followed by concatenation and output projection. Right: Three independent gauge transformations that preserve the function: query-key transformations preserve attention scores through inverse-transpose coupling, value-output transformations preserve head contributions through inverse coupling, and head permutations exploit exchangeability.

Table 1: Assumptions A1–A9: purpose, where used, and architectural examples

Asm.	Mathematical Statement	Where Used	Architectural Examples
A1	Standard multi-head architecture, no weight sharing	Thm. 8	All production models
A2	$d_{\text{model}} = h \cdot d_v$ (canonical dimensions)	Lem. 19	BERT-Base, GPT-2: $768 = 12 \times 64$
A3	Biases absent or transform covariantly	All symmetry statements	Standard implementations
A4	$W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}$ have full column rank	Prop. 10	Generic stratum Θ_0
A5	LayerNorm in standard Pre-LN or Post-LN blocks	Lem. 27	Common implementations
A6	$\mathrm{rank}([W_Q^{(i)} \mid W_K^{(i)}]) = 2d_k$, $d_{\text{model}} \geq 2d_k$	Lem. 19	GPT-2: $768 \geq 128 = 2 \times 64$
A7	Linear independence of bilinear forms $\{W_Q^{(i)}(W_K^{(i)})^\top\}_{i=1}^h$	Prop. 10, Lem. 19	Generic Θ_0 (Zariski-open)
A8	W_O has full **row** rank (square and invertible if A2 holds)	Lem. 23, Thm. 22	Generic Θ_0
A9	FFN weights generic; activation non-polynomial (e.g., GELU)	Lem. 7, Thm. 8	Standard models

and A8 (full row rank of W_O), this establishes attention weight identifiability (Lemma 19) and head-wise factorization (Theorem 22). When biases are present, Assumption A3 requires

they transform covariantly under gauge actions; the exact transformation laws are established in Proposition 12.

2.3. Position-Encoded Architectures (RoPE)

For architectures with rotary position embeddings, position-dependent rotations are applied after the linear projections at the token level. For tokens at positions m and n,

$$Q^{\mathrm{RoPE}}(m) = Q\,R(m), \quad K^{\mathrm{RoPE}}(n) = K\,R(n), \tag{2.4}$$

where $R(m) \in \mathrm{SO}(d_k)$ is block-diagonal with 2×2 rotation blocks of angles $m\theta_j$. The attention score depends only on the *relative* rotation:

$$\ell_{mn} = \frac{1}{\sqrt{d_k}} q_m R(m)(k_n R(n))^\top = \frac{1}{\sqrt{d_k}} q_m R(m-n) k_n^\top, \tag{2.5}$$

since $R(m)R(n)^\top = R(m-n)$. Any gauge transform $A \in \mathrm{GL}(d_k)$ that preserves attention must commute with all $R(m)$, hence

$$\mathcal{C}_{\mathrm{RoPE}} = \{A \in \mathrm{GL}(d_k) \mid AR(m) = R(m)A \text{ for all } m\}. \tag{2.6}$$

Standard RoPE uses distinct frequencies θ_j; when multiple planes share a frequency (non-generic), the commutant enlarges on that collision block to $\mathrm{GL}(r, \mathbb{C})$.

For standard RoPE with 2×2 rotation blocks and distinct frequencies, the commutant consists of block-diagonal matrices where each block has form $aI_2 + bJ$ with J the $90°$ rotation. This yields $\mathcal{C}_{\mathrm{RoPE}} \cong (\mathrm{GL}(1, \mathbb{C}))^{d_k/2}$ with real dimension d_k.

3. Global Maximality of the Gauge Group

We now establish the complete gauge group structure for the canonical Transformer family. Our proof strategy proceeds through three stages: constructing the gauge group candidate, proving these transformations preserve the function, and establishing no additional symmetries exist.

Definition 1 (Standard Gauge Transformations) *The standard gauge group $G_{\max}$ consists of transformations parametrized by $(A_i, C_i) \in \mathrm{GL}(d_k) \times \mathrm{GL}(d_v)$ for each head $i \in \{1, \ldots, h\}$ and permutations $\sigma \in S_h$, acting as:*

$$(W_Q^{(i)},\, W_K^{(i)}) \mapsto \big(W_Q^{(\sigma(i))}\, A_i,\; W_K^{(\sigma(i))}\, (A_i^{-1})^\top\big), \tag{3.1}$$

$$(W_V^{(i)},\, W_{O,i}) \mapsto \big(W_V^{(\sigma(i))}\, C_i,\; C_i^{-1}\, W_{O,\sigma(i)}\big). \tag{3.2}$$

We index (A_i, C_i) by the target *head i. With this action the composition law coincides with the standard wreath product $((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$.*

Theorem 2 (Global Maximality on the Generic Stratum) *For the canonical Transformer family satisfying assumptions A1–A4, A6–A8, the gauge group on the generic stratum Θ_0 equals exactly $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$. No additional parameter symmetries exist beyond those in this group.*

Remark 3 (Role of A7 and A8) *Assumption A7 (linear independence of the bilinear forms $\{W_Q^{(i)}(W_K^{(i)})^\top\}_{i=1}^h$) implies analytic independence of attention weights via Proposition 10. Assumption A8 requires full row rank of W_O, ensuring the value-output sector has no extra degeneracies.*

Proof sketch. We establish equality via bidirectional containment. The forward direction $G_{\max} \subseteq G(\theta)$ is by direct verification.

For the reverse $G(\theta) \subseteq G_{\max}$: (i) **Lie algebra.** Preserving $Q_i K_i^\top$ and $V_i W_{O,i}$ forces $\delta W_Q^{(i)} = W_Q^{(i)} X_i$, $\delta W_K^{(i)} = -W_K^{(i)} X_i^\top$ and $\delta W_V^{(i)} = W_V^{(i)} Y_i$, $\delta W_{O,i} = -Y_i W_{O,i}$, hence $\mathfrak{g}_{\max} = \bigoplus_{i=1}^h \mathfrak{gl}(d_k) \oplus \bigoplus_{i=1}^h \mathfrak{gl}(d_v)$. (ii) **Identifiability.** By analytic independence of attention weight families (Proposition 10) and uniqueness of analytic mixture representations, attention weights match up to a head permutation (Lemma 19). (iii) **Factorization.** For each head, preservation of $V_i W_{O,i}$ and $Q_i K_i^\top$ yields equal minimal-rank factorizations. By uniqueness of minimal-rank factorizations under A4 and A8, there exist unique $C_i \in \mathrm{GL}(d_v)$ and $A_i \in \mathrm{GL}(d_k)$ with $W_V'^{(i)} = W_V^{(i)} C_i$, $W'_{O,i} = C_i^{-1} W_{O,i}$ and $W_Q'^{(i)} = W_Q^{(i)} A_i$, $W_K'^{(i)} = W_K^{(i)} (A_i^{-1})^\top$ (Theorem 22). (iv) **No cross-head mixing.** By the block-diagonality argument (Lemma 23), transformations cannot mix parameters across heads except via permutations. In the canonical case $d_{\text{model}} = h\, d_v$, this forces $W_O = P W_O$ with P a block permutation, hence $P = I$ after reindexing.

Complete proofs appear in Appendices B.1–B.6. □

Corollary 4 (Gauge Dimension) *The continuous gauge group dimension for canonical Transformers is $h(d_k^2 + d_v^2)$ per layer. For standard configurations with $h = 12$ heads and $d_k = d_v = 64$, this yields $12 \times (64^2 + 64^2) = 98,304$ continuous degrees of freedom per attention layer.*

Corollary 5 (Gauge Group for RoPE Architectures) *For architectures with rotary position embeddings, the gauge group becomes $G_{\text{RoPE}} = ((\mathcal{C}_{\text{RoPE}})^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$ where $\mathcal{C}_{\text{RoPE}}$ is the commutant of position rotations. For standard RoPE with 2×2 rotation blocks, $\mathcal{C}_{\text{RoPE}} \cong (\mathrm{GL}(1, \mathbb{C}))^{d_k/2}$ has real dimension d_k, reducing gauge freedom from d_k^2 to d_k per head in the query-key sector.*

4. Architectural Extensions and Multi-Layer Structure

Having established global maximality for multi-head attention, we extend our analysis to complete Transformer architectures. We prove that LayerNorm preserves gauge symmetries, characterize the combined attention-FFN block structure, and establish that multi-layer Transformers have a direct product gauge group with no inter-layer coupling.

4.1. LayerNorm Preserves Gauge Symmetry

Corollary 6 (LayerNorm compatibility) *Since MHA's output is gauge-invariant and LayerNorm acts on that tensor (Post-LN: $X + \mathrm{MHA}(X)$; Pre-LN: $\mathrm{LN}(X)$), LayerNorm preserves the symmetry in both variants. See Appendix C for proof.*

4.2. Feed-Forward Networks and Block Structure

Feed-forward networks in Transformer blocks possess limited symmetry structure due to non-linear activations.

Lemma 7 (FFN Gauge Structure) *The gauge group of the feed-forward network with GELU activation consists only of hidden unit permutations: $G_{FFN} = S_{d_{ff}}$ where d_{ff} is the hidden dimension.*

Since MHA and FFN operate on disjoint parameters and residual connections prevent coupling, their gauge groups combine as a direct product: $G_{\text{Block}} = G_{\text{MHA}} \times G_{\text{FFN}}$.

4.3. Multi-Layer Direct Product Structure

Theorem 8 (Layer-Local Product on Generic Stratum) *For any parameters, $G_{Model} \subseteq \prod_{\ell=1}^{L} G_{Block}^{(\ell)}$. If assumptions A1–A6 hold (including no parameter tying across layers, Assumption A1) and parameters are in generic position, then equality holds:*

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

Proof [Proof Sketch] Inclusion always holds by construction. For equality, residual connections force any transformation on the block output at layer ℓ to apply identically to the untouched residual stream, which is impossible unless the transformation is the identity. By Lemma 27, LayerNorm is not equivariant under general linear scalings $M \neq I$ on the generic stratum. Consequently, a layer's gauge action cannot propagate across layers, yielding the layerwise direct product. Complete proof in Appendix C. ■

5. Implications for Optimization and Loss Landscape Geometry

The gauge structure has direct consequences for optimization geometry, creating flat directions in the loss landscape and explaining empirical phenomena in Transformer training.

5.1. Hessian Nullspace Structure

Proposition 9 (Hessian Nullspace from Gauge Symmetry) *At any critical point $\theta^* \in \Theta_0$, the Hessian $\nabla^2 L(\theta^*)$ has nullspace dimension at least $h(d_k^2 + d_v^2)$ per layer, with null directions corresponding to the Lie algebra $\mathfrak{g}_{\max}$ of gauge transformations.*

Since gauge transformations preserve the network function, they preserve the loss: $L(g(\theta)) = L(\theta)$ for all $g \in G_{\max}$. Along any one-parameter subgroup $g_t = \exp(tX)$ where $X \in \mathfrak{g}_{\max}$, the loss remains constant, yielding zero curvature in these directions. For an L-layer model, the nullspace dimension is at least $L \cdot h(d_k^2 + d_v^2)$.

5.2. Optimization in Quotient Space

The gradient $\nabla L(\theta)$ is orthogonal to gauge orbits at every point, so gradient descent naturally respects the gauge without explicit constraints. Optimization effectively proceeds on the quotient $\Theta_0/G_{\max}$ of gauge-inequivalent configurations; its effective dimension is $\dim_{\text{eff}} = \dim(\Theta) - L\,h(d_k^2 + d_v^2)$.

This dimension reduction partially explains why Transformers with hundreds of millions of parameters can be trained effectively—the true degrees of freedom are substantially fewer than the raw parameter count.

5.3. Mode Connectivity and Flat Minima

Critical points of Transformer training form continuous manifolds rather than isolated points. Within each connected component of the gauge orbit, all points represent identical functions with identical loss. This provides a geometric explanation for the prevalence of flat minima

observed empirically and the phenomenon of linear mode connectivity between independently trained models. Two parameter configurations that differ by an element of the identity component $G_{\max}^0$ are connected by a flat path in the loss landscape with constant loss value; parameters related only by elements in the discrete components of $G_{\max}$ (e.g., permutations) are not connected by such gauge-induced continuous paths.

The gauge structure also explains why standard sharpness measures based on Hessian eigenvalues are problematic for Transformers. The Hessian necessarily has at least $L{\cdot}h(d_k^2{+}d_v^2)$ zero eigenvalues from gauge directions. Meaningful sharpness must be measured on the quotient space or restricted to gauge-orthogonal directions to avoid including these artificial flat directions.

Takeaways. In summary, gauge symmetries imply that gradient-based training is effectively carried out on the quotient space $\Theta_0/G_{\max}$, since parameters that differ by a gauge transformation define the same model. At well-trained solutions this yields at least $L\,h(d_k^2 + d_v^2)$ exact zero eigenvalues in the Hessian, corresponding to flat directions along gauge orbits. As a consequence, conventional sharpness and mode-connectivity measures can be dominated by these directions, so meaningful notions of sharpness should either quotient out the gauge group or restrict attention to gauge-orthogonal perturbations.

6. Experimental Validation

We validate our theoretical predictions on pretrained GPT-2 checkpoints obtained from Hugging Face, spanning GPT-2 (124M parameters), GPT-2-Medium (355M), GPT-2-Large (774M), and GPT-2-XL (1.5B). All experiments use single-precision (FP32) arithmetic with deterministic settings: fixed random seeds, dropout disabled, and evaluation mode for all layers. For each model, we extract per-head projection matrices (W_Q, W_K, W_V, W_O) and corresponding biases, apply gauge transformations drawn from $G_{\max}$ with well-conditioned matrices, and measure relative errors between original and transformed attention outputs and logits. Full implementation details and extended results appear in Appendix F.

For valid gauge transformations, attention and logit outputs agree up to floating-point roundoff. Across all four GPT-2 models, relative attention-output errors are tightly concentrated around 10^{-6}: medians are approximately 1.2×10^{-6}, with maxima below 3×10^{-6} (tens of $\varepsilon_{\text{mach}}^{\text{FP32}}$). Logit-level relative errors remain below 3.1×10^{-5} (a few hundred $\varepsilon_{\text{mach}}^{\text{FP32}}$) across all checkpoints. As a stricter end-to-end test, we apply the KV gauge-fixing transformation given by Algorithm 1 and compare generation behavior before and after: for every GPT-2 size, greedy decoding produces *identical* token sequences, confirming functional equivalence at FP32 precision.

These experiments operate on full multi-head attention blocks and are consistent with the direct-product structure

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

in which query–key and value–output sectors transform independently and heads may be permuted without changing the attention map. The observed invariance under arbitrary per-head gauges implies invariance of each individual sector when the others are held fixed.

To probe maximality, we construct families of deliberately invalid transformations that violate the structure of $G_{\max}$: (i) asymmetric query–key transforms using independent matrices for W_Q and W_K, (ii) transforms that apply A instead of $A^{-\top}$ to the key sector,

and (iii) value transformations without the compensating inverse on W_O. On all four GPT-2 checkpoints, these invalid modifications produce relative logit errors between 10^1 and 10^4, and the resulting generations diverge completely from the original model. The sharp gap between $O(10^{-6})$–$O(10^{-5})$ errors for valid transformations and $O(10^1)$–$O(10^4)$ errors for invalid ones provides strong empirical evidence that $G_{\max}$ captures exactly the symmetries of the attention mechanism in pretrained Transformers.

7. Practical Implications and Applications

The gauge structure enables several practical applications in model compression, optimization, and analysis.

7.1. Quantifying Parameter Redundancy

A standard Transformer Base model with 110M parameters contains approximately 1.18M continuously redundant dimensions from gauge symmetries, representing 1.1% of total parameters. While modest as a percentage, this corresponds to over one million parameters that can be varied without changing the function, fundamentally affecting optimization geometry. For larger models, the redundancy scales proportionally. GPT-3 scale models with 175B parameters contain over 300M redundant dimensions. The redundancy ratio $h(d_k^2 + d_v^2)/d_{\text{total}}$ decreases with model scale when head dimensions remain fixed while the number of heads increases, as is common in production architectures. This scaling property suggests that architectural innovations reducing gauge redundancy while preserving expressiveness could lead to more parameter-efficient models.

7.2. Model Compression via Gauge-Fixing

Gauge-fixing to a deterministic canonical form enables exact, lossless compression of a large fraction of the continuous gauge redundancy. In the value–output sector we fix $(W_V^{(i)})^\top W_V^{(i)} = I_{d_v}$ and absorb the corresponding change into $W_{O,i}$, which reduces the $\mathrm{GL}(d_v)$ freedom to a residual orthogonal group $O(d_v)$; this removes $\frac{d_v(d_v+1)}{2}$ continuous degrees of freedom per head and leaves an unavoidable $\frac{d_v(d_v-1)}{2}$ residual. In the query–key sector we balance the Grams so that $(W_Q^{(i)})^\top W_Q^{(i)} = (W_K^{(i)})^\top W_K^{(i)}$, which similarly reduces the $\mathrm{GL}(d_k)$ freedom to a residual orthogonal $O(d_k)$, removing $\frac{d_k(d_k+1)}{2}$ degrees per head and leaving a residual $\frac{d_k(d_k-1)}{2}$. Overall, per layer, the deterministic slice eliminates

$$h\Big(\tfrac{d_v(d_v+1)}{2} + \tfrac{d_k(d_k+1)}{2}\Big)$$

continuous gauge degrees of freedom, with the remaining redundancy exactly $O(d_k) \times O(d_v)$ per head. We *default* to a numerically stable QR-based canonical slice (Algorithm 1); an alternative Gram-balanced approach (setting $(W_Q^{(i)})^\top W_Q^{(i)} = (W_K^{(i)})^\top W_K^{(i)}$) achieves the same reduction and is discussed in Appendix H.

Query–key canonical form (QR). For each head i, compute the thin QR factorization $W_Q^{(i)} = Q_Q R_Q$ with $\mathrm{diag}(R_Q) > 0$ (flip signs if needed). Apply the gauge $A_i := R_Q^{-1}$:

$$W_Q^{(i)} \leftarrow W_Q^{(i)} A_i = Q_Q, \qquad W_K^{(i)} \leftarrow W_K^{(i)} (A_i^{-1})^\top = W_K^{(i)} R_Q^\top.$$

This enforces $(W_Q^{(i)})^\top W_Q^{(i)} = I_{d_k}$ and reduces the $\mathrm{GL}(d_k)$ freedom to a *residual* $O(d_k)$, which we leave unfixed in our canonical representative. (Any further orthogonal tie-breaker would be optional and is not needed for our theorems or compression counts.)

Value–output canonical form (QR). Compute the thin QR factorization $W_V^{(i)} = Q_V R_V$ with $\mathrm{diag}(R_V) > 0$ and set $C_i := R_V^{-1}$:

$$W_V^{(i)} \leftarrow W_V^{(i)} C_i = Q_V, \qquad W_{O,i} \leftarrow C_i^{-1} W_{O,i} = R_V\, W_{O,i}.$$

This enforces $(W_V^{(i)})^\top W_V^{(i)} = I_{d_v}$ and reduces the $\mathrm{GL}(d_v)$ freedom to a *residual* $O(d_v)$, which we likewise leave unfixed in our canonical representative. (Again, any orthogonal tie-breaker would be optional and is not used in our results.)

Head ordering (break S_h). Sort the heads by $\|W_K^{(i)}\|_F$ in descending order to fix the permutation freedom. If two heads happen to have exactly the same norm (a nongeneric, measure-zero event), break ties using any fixed deterministic rule (e.g., a lexicographic order on additional invariants).

This compression is exact: the network function is preserved identically while removing $h\big(\frac{d_v(d_v+1)}{2} + \frac{d_k(d_k+1)}{2}\big)$ continuous degrees of freedom per layer, with a residual $h\big(\frac{d_v(d_v-1)}{2} + \frac{d_k(d_k-1)}{2}\big)$ orthogonal freedom remaining in the value–output and query–key sectors. The per-layer cost is $\mathcal{O}\big(h(d_{\text{model}} d_k^2 + d_{\text{model}} d_v^2)\big)$ for the factorization steps. Implementation details and the full procedure appear in Algorithm 1 in the appendix.

7.3. Gauge-Aware Optimization

Standard optimizers waste computation updating along gauge directions where the function remains unchanged. Projecting gradients onto gauge-orthogonal subspaces eliminates these redundant updates, potentially accelerating convergence. The projection can be implemented efficiently using the structure of $\mathfrak{g}_{\max}$:

$$\nabla_{\text{proj}} L = \nabla L - \Pi_{\mathfrak{g}} \nabla L \tag{7.1}$$

where $\Pi_{\mathfrak{g}}$ denotes projection onto the gauge tangent space. This adds negligible computational overhead while focusing optimization on function-changing directions.

7.4. Model Merging and Averaging

Independent training yields different gauge-equivalent representations even when models converge to similar functions. Naive parameter averaging fails because models occupy different points in the same gauge orbit. Gauge alignment before averaging enables meaningful parameter interpolation by transforming models to a common gauge. The alignment procedure involves solving Procrustes problems for optimal (A_i, C_i) per head and finding the optimal head permutation via the Hungarian algorithm. This explains why sophisticated model merging techniques succeed where simple averaging fails, and provides a principled framework for developing improved merging algorithms.

7.5. Implications for Architecture Design

The gauge structure provides insights for architectural innovations. Multi-query attention and grouped-query attention deliberately couple heads by sharing key-value projections, reducing gauge freedom at the cost of expressiveness. This trade-off can now be quantified

precisely through the dimension of the resulting gauge group. Future architectures might exploit gauge structure more deliberately, perhaps by operating directly on gauge-invariant features or incorporating gauge-aware regularization during training.

8. Related Work

Parameter symmetries in neural networks have been studied extensively. For fully connected networks, permutation symmetries arising from neuron reordering create S_n symmetry groups [9; 13]. These discrete symmetries are substantially smaller than the continuous gauge groups we identify in Transformers. Convolutional networks exhibit translation invariance and filter permutation symmetries [14; 3].

For Transformers specifically, recent work has begun identifying partial symmetries. van Nierop [17] identified gauge invariance from a physics perspective without proving completeness. Entezari et al. [7] studied permutation invariance for mode connectivity. da Silva et al. [4]; Zhang et al. [19] demonstrated symmetry effects on sharpness and model fusion. Henry et al. [10] analyzed simplified attention without softmax, which exhibits fundamentally different geometry.

Previous optimization work recognized symmetry-induced flat directions [12] and mode connectivity [8; 6]. Model merging techniques [1; 18] implicitly leverage symmetries through weight matching. Our complete characterization with maximality proof provides the theoretical foundation for these observations, explaining why certain techniques succeed and suggesting systematic improvements.

9. Discussion and Future Directions

Our characterization opens several research directions. Gauge-aware optimization algorithms that explicitly project out redundant directions could accelerate training. Theoretical analysis of how gauge structure affects generalization may yield new complexity measures beyond parameter counting. Understanding how architectural innovations like mixture-of-experts or state-space models modify gauge structure could guide design choices. The reduction in gauge freedom from RoPE architectures suggests a connection between symmetry constraints and parameter efficiency. Architectures that deliberately break certain symmetries while preserving others might achieve better trade-offs between expressiveness and efficiency. The gauge perspective also suggests new initialization schemes that distribute parameters uniformly across gauge orbits rather than in raw parameter space.

10. Conclusion

We have established the complete gauge group structure of Transformer architectures, proving global maximality for the canonical family under mild generic conditions. The gauge group $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$ captures all parameter redundancies in standard multi-head attention, with no additional symmetries beyond those we identify. This definitive characterization resolves the fundamental question of parameter-function correspondence for Transformers.

Our proof of maximality through three independent mathematical arguments—Lie algebra characterization, attention weight identifiability, and necessary factorization—ensures completeness. These results extend through LayerNorm and residual connections, yielding the direct product structure $G_{\text{Model}} = \prod_{l=1}^{L} G_{\text{Block}}^{(l)}$ for multi-layer architectures.

The gauge structure reveals substantial parameter redundancy exceeding one million continuous degrees of freedom in standard models. These redundant directions correspond to flat regions in the loss landscape, explaining empirical phenomena and enabling practical advances in compression, optimization, and model merging. Our experiments on pretrained GPT-2 models confirm that these symmetries are not merely theoretical artifacts but manifest precisely in production-scale architectures. As Transformers continue driving advances in artificial intelligence, leveraging their mathematical structure becomes increasingly critical for both theoretical understanding and practical improvements.

References

[1] Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CQsmMYmlP5T.

[2] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37 (6A):3099–3132, 2009. doi: 10.1214/09-AOS689. URL https://doi.org/10.1214/09-AOS689.

[3] Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999. PMLR, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

[4] Marvin F. da Silva, Felix Dangel, and Sageev Oore. Hide & seek: Transformer symmetries obscure sharpness & riemannian geometry finds it. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, volume 267 of *Proceedings of Machine Learning Research*, pages 55591–55613. PMLR, 2025. URL https://icml.cc/virtual/2025/poster/46431. Spotlight Poster.

[5] 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. URL https://arxiv.org/abs/1810.04805.

[6] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1309–1318. PMLR, 2018. URL https://proceedings.mlr.press/v80/draxler18a.html.

[7] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dNigytemkL.

[8] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew G. Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in*

Neural Information Processing Systems, pages 8803–8812, 2018. URL https://arxiv.org/abs/1802.10026.

[9] Robert Hecht-Nielsen. On the algebraic structure of feedforward network weight spaces. *Advanced Neural Computers*, pages 129–135, 1990. URL https://doi.org/10.1016/B978-0-444-88400-8.50019-4.

[10] Nathan W. Henry, Giovanni Luca Marchetti, and Kathlén Kohn. Geometry of lightning self-attention: Identifiability and dimension. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=XtY3xYQWcW. Poster.

[11] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, pages 876–885. AUAI Press, 2018. URL https://arxiv.org/abs/1803.05407.

[12] Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel L.K. Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. *arXiv preprint arXiv:2012.04728*, 2020. URL https://arxiv.org/abs/2012.04728.

[13] Věra Kurková and Paul C. Kainen. Functionally equivalent feedforward neural networks. *Neural Computation*, 6(3):543–558, 1994. URL https://doi.org/10.1162/neco.1994.6.3.543.

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. URL https://doi.org/10.1109/5.726791.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Technical report.

[16] Henry Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269, 1963. doi: 10.1214/aoms/1177703862. URL https://doi.org/10.1214/aoms/1177703862.

[17] Leo van Nierop. Transformer models are gauge invariant: A mathematical connection between AI and particle physics, 2024. URL https://arxiv.org/abs/2412.14543.

[18] Mitchell Wortsman et al. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

[19] Binchi Zhang, Zaiyi Zheng, Zhengzhang Chen, and Jundong Li. Beyond the permutation symmetry of transformers: The role of rotation for model fusion. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, volume 267 of *Proceedings of Machine Learning Research*, pages 77090–77106. PMLR, 2025. URL https://icml.cc/virtual/2025/poster/43634. Spotlight Poster.

Appendix A. Mathematical Framework and Complete Assumptions

This appendix provides complete justifications for all assumptions and establishes the mathematical framework underlying our gauge group characterization.

A.1. Complete Statements of Assumptions A1–A9

For convenience we restate here the structural and genericity assumptions used throughout the paper. They coincide with the entries of Table 1 in the main text.

A1 **Standard multi-head architecture, no weight sharing.** Each layer uses independent parameters $\{W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}, W_{O,i}\}_{i=1}^h$; there is no parameter sharing across heads or layers.

A2 **Canonical dimensions.** The model dimension satisfies $d_{\text{model}} = h\, d_v$, so that the concatenated value map $V \in \mathbb{R}^{n \times h d_v}$ and the output projection $W_O \in \mathbb{R}^{(h d_v) \times d_{\text{model}}}$ have compatible shapes in the canonical case.

A3 **Biases absent or transform covariantly.** Either biases are omitted, or whenever gauge transformations are applied we transform all bias vectors covariantly so that the network function is preserved (cf. Proposition 12).

A4 **Full column rank of per-head projections.** For each head i, the matrices $W_Q^{(i)} \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_K^{(i)} \in \mathbb{R}^{d_{\text{model}} \times d_k}$ and $W_V^{(i)} \in \mathbb{R}^{d_{\text{model}} \times d_v}$ have full column rank.

A5 **Standard LayerNorm blocks.** LayerNorm is applied in either the standard Pre-LN or Post-LN configuration, with learned affine parameters (γ, β) that are not constrained or tied across layers.

A6 **Sufficient width for attention identifiability.** For each head i we have $\operatorname{rank}\big([W_Q^{(i)} \mid W_K^{(i)}]\big) = 2d_k$ and $d_{\text{model}} \geq 2d_k$, ensuring that the bilinear forms $X W_Q^{(i)} (W_K^{(i)})^\top X^\top$ can be separated.

A7 **Linear independence of query–key bilinear forms.** The matrices $\{W_Q^{(i)} (W_K^{(i)})^\top\}_{i=1}^h$ are linearly independent as elements of $\mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ on the generic stratum Θ_0.

A8 **Full row rank of the output projection.** The output matrix $W_O \in \mathbb{R}^{(h d_v) \times d_{\text{model}}}$ has full row rank (and is square and invertible in the canonical case $d_{\text{model}} = h d_v$).

A9 **Generic feed-forward blocks.** Feed-forward weight matrices are in generic position and the nonlinearity is non-polynomial (e.g. GELU, ReLU), ruling out accidental additional symmetries in the FFN.

A.2. Linear Independence of Attention Weight Maps

Proposition 10 (Analytic Independence of Attention Weights) *Assume A1–A9 and fix a sequence length n. For each head $i \in \{1, \dots, h\}$, the row-wise attention map*

$$\alpha_i(X) = \mathrm{softmax}\Big(S_i(X)\Big) \in \mathbb{R}^{n\times n}, \qquad S_i(X) = \tfrac{1}{\sqrt{d_k}} X W_Q^{(i)} \big(W_K^{(i)}\big)^\top X^\top,$$

is real-analytic in X. On a Zariski-open dense set of parameters (the generic stratum), the family $\{\alpha_i(\cdot)\}_{i=1}^h$ is linearly independent as functions: if $\sum_{i=1}^h c_i \alpha_i(X) \equiv 0$ for all X, then $c_i = 0$ for all i.

Proof Each $S_i(X)$ is quadratic in X, hence analytic; row-wise softmax is analytic, so α_i is analytic. Linearize at $S = 0$ for the *row-wise* softmax acting on an $n \times n$ matrix S:

$$\mathrm{softmax}(S) = \tfrac{1}{n} J_n + \tfrac{1}{n} S P_n + O(\|S\|^2),$$

where $J_n = \mathbf{1}\mathbf{1}^\top$ and $P_n = I_n - \frac{1}{n} J_n$ is the centering projector. Applying this to $S_i(X)$ yields

$$\alpha_i(X) = \tfrac{1}{n} J_n + \tfrac{1}{n} S_i(X) P_n + O(\|X\|^4).$$

Assume $\sum_i c_i \alpha_i(X) \equiv 0$. The constant term gives $\sum_i c_i = 0$. The quadratic term gives

$$\Big(\sum_i c_i S_i(X)\Big) P_n = 0 \quad \text{for all } X.$$

Write $S(X) = \sum_i c_i S_i(X)$ and note $S(X)P_n = 0$ if and only if every row of $S(X)$ is constant. Since $S(X)_{jk} = \frac{1}{\sqrt{d_k}} X_j M X_k^\top$ with $M := \sum_i c_i W_Q^{(i)} (W_K^{(i)})^\top$ (a bilinear form; symmetry not required), row constancy ($S(X)P_n = 0$) means $S_{jk} = S_{j\ell}$ for all j, k, ℓ, i.e.

$$X_j M X_k^\top = X_j M X_\ell^\top \;\Rightarrow\; X_j M (X_k - X_\ell)^\top = 0.$$

As X_j and $(X_k - X_\ell)$ can independently span $\mathbb{R}^{d_{\text{model}}}$ (as X varies), this forces $M = 0$. On the generic stratum (Assumption A7), the family of bilinear forms $\{W_Q^{(i)} (W_K^{(i)})^\top\}_{i=1}^h$ is linearly independent, hence $M = 0$ implies $c_i = 0$ for all i. ∎

Remark 11 (Constructive Variant) *If each $W_{O,i}$ has full row rank with a right inverse and $\dim \bigcap_{j\neq i} \mathrm{Null}(W_V^{(j)}) \geq d_v$, one can recover α_i by stacked probe inputs; we use the analytic route to avoid ε-scaling while retaining rigor.*

A.3. Bias Transformations and Affine Projections

Proposition 12 (Covariant Bias Transformation) *Consider multi-head attention under the row-vector convention with optional affine biases*

$$Q_i = X W_Q^{(i)} + \mathbf{1}\, b_Q^{(i)\top}, \quad K_i = X W_K^{(i)} + \mathbf{1}\, b_K^{(i)\top}, \quad V_i = X W_V^{(i)} + \mathbf{1}\, b_V^{(i)\top},$$

and a single output bias $b_O \in \mathbb{R}^{d_{model}}$ added after projection of concatenated heads. Assume the standard gauge action in Definition 1:

$$(W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}, W_{O,i}) \;\mapsto\; \big(W_Q^{(i)} A_i,\; W_K^{(i)} (A_i^{-1})^\top,\; W_V^{(i)} C_i,\; C_i^{-1} W_{O,i}\big),$$

with $A_i \in \mathrm{GL}(d_k)$, $C_i \in \mathrm{GL}(d_v)$, and keep the global output bias b_O in the usual place, i.e. $Y = \sum_{i=1}^{h} \alpha_i(X)\, V_i W_{O,i} + \mathbf{1}\, b_O^\top$. Then invariance of the MHA function for all inputs X forces the biases to transform as

$$b_Q'^{(i)} = b_Q^{(i)} A_i, \qquad b_K'^{(i)} = b_K^{(i)} (A_i^{-1})^\top, \qquad b_V'^{(i)} = b_V^{(i)} C_i, \qquad b_O' = b_O.$$

Conversely, with these bias laws, MHA is invariant under the gauge action. On the generic stratum (Assumptions A2, A4, A8), these transformations are unique.

Proof

Query–key sector. Requiring equality of score bilinears for all X,

$$\big(XW_Q^{(i)} A_i + \mathbf{1}\, b_Q'^{(i)\top}\big)\big(XW_K^{(i)} (A_i^{-1})^\top + \mathbf{1}\, b_K'^{(i)\top}\big)^\top \;=\; \big(XW_Q^{(i)} + \mathbf{1}\, b_Q^{(i)\top}\big)\big(XW_K^{(i)} + \mathbf{1}\, b_K^{(i)\top}\big)^\top,$$

and matching the terms linear in X (the identity must hold for all X) gives

$$A_i\, b_K'^{(i)\top} = b_K^{(i)\top}, \qquad b_Q'^{(i)\top} (A_i^{-1}) = b_Q^{(i)\top},$$

hence

$$b_K'^{(i)} = b_K^{(i)} (A_i^{-1})^\top, \qquad b_Q'^{(i)} = b_Q^{(i)} A_i.$$

The quadratic terms cancel because $A_i(A_i^{-1}) = I$, and the remaining constant term is then satisfied. Uniqueness on the generic stratum follows by the same coefficient matching and the full-column-rank assumption A4.

Value–output sector. Write the (per-head) affine value map and the global output:

$$V_i = XW_V^{(i)} + \mathbf{1}\, b_V^{(i)\top}, \qquad Y \;=\; \sum_{i=1}^{h} \alpha_i(X)\, V_i W_{O,i} \;+\; \mathbf{1}\, b_O^\top.$$

Under the gauge action,

$$(W_V^{(i)}, W_{O,i}) \mapsto \big(W_V^{(i)} C_i,\; C_i^{-1} W_{O,i}\big), \qquad V_i' = XW_V^{(i)} C_i + \mathbf{1}\, b_V'^{(i)\top},$$

and the transformed output is

$$Y' \;=\; \sum_{i=1}^{h} \alpha_i(X)\, \big(XW_V^{(i)} C_i + \mathbf{1}\, b_V'^{(i)\top}\big)\, C_i^{-1} W_{O,i} \;+\; \mathbf{1}\, b_O'^{\top}.$$

Since $\alpha_i(X)$ are row-stochastic, $\alpha_i(X)\mathbf{1} = \mathbf{1}$ for all i and X. Thus the bias contribution is X-independent, and equality $Y' = Y$ for all X gives

$$\sum_{i=1}^{h} b_V'^{(i)\top} C_i^{-1} W_{O,i} \;+\; b_O'^{\top} \;=\; \sum_{i=1}^{h} b_V^{(i)\top} W_{O,i} \;+\; b_O^\top. \tag{A.1}$$

Define row vectors

$$u' := \big[b_V'^{(1)\top} C_1^{-1} \;\big|\; \cdots \;\big|\; b_V'^{(h)\top} C_h^{-1}\big] \in \mathbb{R}^{1\times(hd_v)}, \quad u := \big[b_V^{(1)\top} \;\big|\; \cdots \;\big|\; b_V^{(h)\top}\big] \in \mathbb{R}^{1\times(hd_v)}.$$

With the standard block stacking $W_O = \begin{bmatrix} W_{O,1}^\top & \cdots & W_{O,h}^\top \end{bmatrix}^\top \in \mathbb{R}^{(hd_v)\times d_{\text{model}}}$, (A.1) is

$$u' W_O \;+\; b_O'^\top \;=\; u W_O \;+\; b_O^\top. \tag{A.2}$$

Uniqueness and the role of b_O. The global output bias b_O is added after all head-wise gauge operations and is not transformed by Definition 1; we therefore treat it as gauge-inert and set $b'_O = b_O$. With this convention, (A.2) reduces to $(u' - u)\, W_O = 0$. By A8, W_O has full *row* rank, so its left nullspace is trivial; hence $u' = u$, which yields $b_V'^{(i)} = b_V^{(i)} C_i$ for each head i. (In the canonical case A2, W_O is square and invertible, and the same conclusion follows immediately.)

Sufficiency. Substituting $b_Q'^{(i)} = b_Q^{(i)} A_i$, $b_K'^{(i)} = b_K^{(i)} (A_i^{-1})^\top$, $b_V'^{(i)} = b_V^{(i)} C_i$, and $b'_O = b_O$ into the expanded expressions shows the affine cross-terms cancel and all constants match, so the MHA function is preserved. ∎

Remark 13 (Broadcasting and placement of b_O) *The vectors $b_Q^{(i)}, b_K^{(i)}, b_V^{(i)}$ are broadcast across the sequence length via $\mathbf{1}\, b^\top$. The attention weights satisfy $\alpha_i(X)\mathbf{1} = \mathbf{1}$, so the bias contribution is independent of X. The output bias b_O is a single d_{model}-vector added after concatenation and projection; it is therefore unaffected by head-wise C_i (hence $b'_O = b_O$).*

A.4. Extended Geometric Intuition

Example 1 (Concrete Gauge Transformation with Complete Calculations) *For $W_Q = I_2$ and $W_K = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$, consider the transformation with $A = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$. Setting $W'_Q = W_Q A$ and $W'_K = W_K (A^{-1})^\top$ yields:*

First, compute $A^{-1} = \begin{bmatrix} 1/2 & -1/2 \\ 0 & 1 \end{bmatrix}$ and $(A^{-1})^\top = \begin{bmatrix} 1/2 & 0 \\ -1/2 & 1 \end{bmatrix}$.

Then:

$$W'_Q = I_2 \cdot \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \tag{A.3}$$

$$W'_K = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1/2 & 0 \\ -1/2 & 1 \end{bmatrix} = \begin{bmatrix} 1/2 & 1 \\ -1 & 3 \end{bmatrix} \tag{A.4}$$

Direct computation verifies:

$$W'_Q (W'_K)^\top = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1/2 & -1 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} = W_Q W_K^\top \tag{A.5}$$

The attention scores remain identical despite the parameter transformation.

A.5. Technical Remarks and Contextual Notes

Remark 14 (Architectural Scope) *Global maximality requires $d_{model} = h \cdot d_v$, satisfied by BERT-Base ($768 = 12 \times 64$), GPT-2 ($768 = 12 \times 64$), and GPT-3 ($12,288 = 96 \times 128$). For architectures where this dimensional relationship does not hold exactly, continuous maximality of the gauge group is established, though the discrete structure may differ from the pure permutation group S_h.*

Remark 15 (Generic Stratum and Optimization Trajectories) *The generic stratum Θ_0 where our results hold has full Lebesgue measure in the parameter space. Since optimization trajectories are absolutely continuous with respect to Lebesgue measure, they avoid the measure-zero exceptional set with probability one. This means our characterization applies to essentially all parameter configurations encountered during training, not just at initialization.*

Remark 16 (Essential Nature of Architectural Constraints) *Several architectural choices that appear conventional are actually mathematically essential for gauge symmetry preservation:*

1. *LayerNorm placement after the output projection W_O rather than before is necessary. If LayerNorm operated on concatenated head outputs before projection, the value-output transformations C_i would change individual head statistics, breaking gauge invariance.*

2. *The absence of weight sharing across heads is crucial. Shared parameters would couple gauge transformations across heads, preventing the direct product structure.*

3. *The standard scaling factor $1/\sqrt{d_k}$ stabilizes dot-product magnitudes and gradients. By contrast, gauge invariance itself follows from preserving $QK^\top$ under $(W_Q, W_K) \mapsto (W_Q A, W_K (A^{-1})^\top)$ and does not depend on this scaling.*

Remark 17 (Practical Implications of Gauge Structure) *The extensive flat directions from gauge symmetry help explain several empirical phenomena:*

- *The prevalence of wide minima in Transformer optimization despite high parameter counts*
- *Success of model averaging after gauge alignment [11; 18]*
- *Effectiveness of lottery ticket pruning within gauge equivalence classes*
- *Convergence to similar performance from diverse initializations*

These observations suggest that apparent complexity in parameter space masks geometric simplicity in function space.

Appendix B. Complete Proofs for Gauge Group Maximality

This section provides the full proof of Theorem 2. We first show that every element of the proposed gauge group $G_{\max}$ preserves the multi-head attention map (sufficiency). We then characterize all infinitesimal symmetries at a generic parameter via the Lie algebra of $G_{\max}$ and identifiability of attention maps, and finally upgrade these local statements to a global classification, proving that no additional symmetries exist and that $G_{\max}$ is maximal on the generic stratum Θ_0.

B.1. Proof of Sufficiency

Lemma 18 (Every transformation in $G_{\max}$ preserves MHA) *For all $g \in G_{\max}$, $\theta \in \Theta_0$, and inputs $X \in \mathbb{R}^{n \times d_{model}}$, we have $MHA(X; g(\theta)) = MHA(X; \theta)$.*

Proof Consider transformations with identity permutation first. For each head i, the query-key transformation preserves attention scores:

$$Q_i'(K_i')^\top = XW_Q^{(i)}A_i \cdot (A_i^{-1})^\top (W_K^{(i)})^\top X^\top \tag{B.1}$$
$$= XW_Q^{(i)}A_i(A_i^{-1})^\top (W_K^{(i)})^\top X^\top \tag{B.2}$$
$$= XW_Q^{(i)}I_{d_k}(W_K^{(i)})^\top X^\top \tag{B.3}$$
$$= XW_Q^{(i)}(W_K^{(i)})^\top X^\top = Q_iK_i^\top \tag{B.4}$$

Since softmax operates element-wise on each row, the attention weights $\alpha_i(X) = \text{softmax}(Q_iK_i^\top/\sqrt{d_k})$ remain unchanged.

The value-output transformation preserves each head's contribution:

$$A_i'(X)W_{O,i}' = \alpha_i(X)V_i'W_{O,i}' \tag{B.5}$$
$$= \alpha_i(X)XW_V^{(i)}C_i \cdot C_i^{-1}W_{O,i} \tag{B.6}$$
$$= \alpha_i(X)XW_V^{(i)}W_{O,i} = A_i(X)W_{O,i} \tag{B.7}$$

For permutations $\sigma \in S_h$, we have:

$$\text{MHA}(X; g(\theta)) = \sum_{i=1}^{h} A_{\sigma^{-1}(i)}(X)W_{O,\sigma^{-1}(i)} \tag{B.8}$$
$$= \sum_{j=1}^{h} A_j(X)W_{O,j} = \text{MHA}(X;\theta) \tag{B.9}$$

where we substituted $j = \sigma^{-1}(i)$. ■

B.2. Proof of Attention Weight Identifiability

Lemma 19 (Attention-weight identifiability via analytic independence) *Assume A1–A9. If* $\text{MHA}(X;\theta) = \text{MHA}(X;\theta')$ *for all* X *in a nonempty open set* $U \subseteq \mathbb{R}^{n\times d_{model}}$*, then there exists* $\sigma \in S_h$ *such that, for all* $X \in U$ *and all* i*,*

$$\alpha_i(X;\theta) = \alpha'_{\sigma(i)}(X;\theta') \quad \textit{and} \quad V_i(X;\theta)W_{O,i}(\theta) = V'_{\sigma(i)}(X;\theta')W'_{O,\sigma(i)}(\theta').$$

Proof We have an identity of analytic mixtures on an open set U:

$$\sum_{i=1}^{h} \alpha_i(X;\theta)\, F_i(X;\theta) \;=\; \sum_{j=1}^{h} \alpha'_j(X;\theta')\, F'_j(X;\theta'),$$

where $\alpha_i(X) = \text{softmax}\big(Q_iK_i^\top/\sqrt{d_k}\big)$ with $Q_i = XW_Q^{(i)}$, $K_i = XW_K^{(i)}$, and $F_i(X) = V_iW_{O,i}$ with $V_i = XW_V^{(i)}$.

By Proposition 10 (using A7), the families $\{\alpha_i(\cdot)\}_{i=1}^h$ and $\{\alpha'_j(\cdot)\}_{j=1}^h$ are linearly independent analytic functions on U. On the same generic stratum, the maps $\{F_i(\cdot)\}$ and $\{F'_j(\cdot)\}$

are pairwise distinct by full-rank assumptions (A4, A8), i.e., equality $F_i \equiv F_j$ (or $F'_i \equiv F'_j$) would impose polynomial constraints on parameters and hence is non-generic.

By uniqueness of analytic mixture representations [16; 2], there exists $\sigma \in S_h$ with

$$\alpha_i(\cdot;\theta) = \alpha'_{\sigma(i)}(\cdot;\theta') \quad \text{and} \quad F_i(\cdot;\theta) = F'_{\sigma(i)}(\cdot;\theta') \quad \text{on } U,$$

hence everywhere by the identity theorem for real-analytic functions. ■

B.3. Proof of Lie Algebra Characterization

Lemma 20 (Lie algebra equals $\mathfrak{g}_{\max}$) *For $\theta \in \Theta_0$, the Lie algebra of the gauge group $G(\theta)$ equals*

$$\mathfrak{g}_{\max} \;=\; \bigoplus_{i=1}^{h} \mathfrak{gl}(d_k) \;\oplus\; \bigoplus_{i=1}^{h} \mathfrak{gl}(d_v).$$

Proof We establish $\mathfrak{g}(\theta) = \mathfrak{g}_{\max}$ by bidirectional inclusion.

Forward inclusion $\big(\mathfrak{g}_{\max} \subseteq \mathfrak{g}(\theta)\big)$. By Lemma 18, every element of $G_{\max}$ preserves MHA. Differentiating the one-parameter subgroups $t \mapsto \exp(tX_i)$ and $t \mapsto \exp(tY_i)$ at $t = 0$ yields the generators

$$\delta W_Q^{(i)} = W_Q^{(i)} X_i, \quad \delta W_K^{(i)} = -\,W_K^{(i)} X_i^\top, \qquad \delta W_V^{(i)} = W_V^{(i)} Y_i, \quad \delta W_{O,i} = -\,Y_i W_{O,i},$$

for arbitrary $X_i \in \mathfrak{gl}(d_k)$ and $Y_i \in \mathfrak{gl}(d_v)$. Hence $\mathfrak{g}_{\max}$ is contained in the Lie algebra of symmetries.

Reverse inclusion $\big(\mathfrak{g}(\theta) \subseteq \mathfrak{g}_{\max}\big)$. Let g_t be a smooth one-parameter family of gauge transformations with $g_0 = \mathrm{id}$ and $\mathrm{MHA}(X; g_t(\theta)) = \mathrm{MHA}(X;\theta)$ for all X and all t. By Lemma 19, attention weights are preserved up to a head permutation. Since permutations form a discrete subgroup, any C^1 path g_t is contained in the identity component, so (after a single reindexing) the permutation is constant and equals id for t near 0. Therefore the first-order invariance decomposes *head-wise*:

$$\delta\big(Q_i K_i^\top\big) \;=\; 0 \quad \text{and} \quad \delta\big(V_i W_{O,i}\big) \;=\; 0 \qquad \text{for each head } i.$$

Query–key sector. Differentiating $Q_i(t)K_i(t)^\top = Q_i K_i^\top$ at $t = 0$ gives, for all inputs X,

$$X\,\delta W_Q^{(i)} (W_K^{(i)})^\top X^\top \;+\; X\, W_Q^{(i)} (\delta W_K^{(i)})^\top X^\top \;=\; 0. \tag{B.10}$$

Let

$$B \;:=\; \delta W_Q^{(i)} (W_K^{(i)})^\top \;+\; W_Q^{(i)} (\delta W_K^{(i)})^\top \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}.$$

Then (B.10) reads $XBX^\top = 0$ for *all* X. Fix arbitrary $u, v \in \mathbb{R}^{d_{\text{model}}}$ and take X whose first two rows are $u^\top$ and $v^\top$ (remaining rows zero). The $(1,2)$ entry of $XBX^\top$ is $u^\top B v$, so $u^\top B v = 0$ for all u, v, which forces $B = 0$.

Thus

$$\delta W_Q^{(i)} (W_K^{(i)})^\top \;=\; -\,W_Q^{(i)} (\delta W_K^{(i)})^\top.$$

Post-multiplying by a fixed left inverse of $(W_K^{(i)})^\top$ (e.g., the Moore–Penrose pseudoinverse transpose) and using full column rank (A4) gives

$$\delta W_Q^{(i)} \;=\; W_Q^{(i)} X_i \quad \text{for some } X_i \in \mathfrak{gl}(d_k).$$

Substituting back yields $\delta W_K^{(i)} = -\,W_K^{(i)} X_i^\top$.

Value–output sector. The same argument applied to $\delta\big(V_i W_{O,i}\big) = 0$ shows

$$\delta W_V^{(i)} = W_V^{(i)} Y_i, \qquad \delta W_{O,i} = -\,Y_i W_{O,i} \quad \text{for some } Y_i \in \mathfrak{gl}(d_v).$$

No cross-head or cross-sector couplings appear. Hence every infinitesimal symmetry has the stated block-diagonal form, so $\mathfrak{g}(\theta) \subseteq \mathfrak{g}_{\max}$. ■

Corollary 21 (Identity component) *The connected component of the identity of the gauge group equals*

$$G(\theta)^0 \;=\; \big(\mathrm{GL}(d_k)^0\big)^h \times \big(\mathrm{GL}(d_v)^0\big)^h.$$

Proof The generators from Lemma 20 integrate to the flows

$$W_Q^{(i)}(t) = W_Q^{(i)}(0)\, e^{tX_i}, \quad W_K^{(i)}(t) = W_K^{(i)}(0)\, e^{-tX_i^\top}, \qquad W_V^{(i)}(t) = W_V^{(i)}(0)\, e^{tY_i}, \quad W_{O,i}(t) = e^{-tY_i}\, W_{O,i}(0),$$

which preserve $Q_i K_i^\top$ and $V_i W_{O,i}$ for all t. Since matrix exponentials generate $\mathrm{GL}(d_k)^0$ and $\mathrm{GL}(d_v)^0$ and the flows decouple across heads and sectors, the identity component is the stated direct product. Conversely, any g in the identity component has tangent at the identity in $\mathfrak{g}_{\max}$, so $G(\theta)^0$ is precisely the group obtained by exponentiating $\mathfrak{g}_{\max}$. ■

B.4. Proof of Factorization Theorem

Theorem 22 (Every gauge transformation factorizes) *Every gauge transformation factors as independent query–key and value–output transformations, composed with a head permutation.*

Proof By Lemma 19, there exists $\sigma \in S_h$ such that, after reindexing heads by σ,

$$\sum_{i=1}^{h} \alpha_i(X)\, V_i W_{O,i} \;=\; \sum_{i=1}^{h} \alpha_i(X)\, V_i' W_{O,i}' \quad \text{for all } X.$$

By Proposition 10 (using A7), the family $\{\alpha_i(\cdot)\}_{i=1}^h$ is linearly independent on the generic stratum, hence for each head i,

$$V_i W_{O,i} \;=\; V_i' W_{O,i}' \qquad \text{for all } X.$$

Value–output sector (column-space argument; A4 and A8 only). Recall $V_i = XW_V^{(i)}$ and $V_i' = XW_V^{(i)\prime}$. The identity $V_i W_{O,i} = V_i' W_{O,i}'$ for all X implies the *parameter* identity

$$L_i \;:=\; W_V^{(i)} W_{O,i} \;=\; W_V^{(i)\prime} W_{O,i}'.$$

By A4, $W_V^{(i)}$ has full column rank d_v. By A8, $W_{O,i}$ has full row rank d_v, hence $W_{O,i} : \mathbb{R}^{d_{\text{model}}} \to \mathbb{R}^{d_v}$ is *surjective*. Therefore

$$\mathrm{Im}(L_i) \;=\; \mathrm{Im}\big(W_V^{(i)} W_{O,i}\big) \;=\; \mathrm{Im}\big(W_V^{(i)}\big),$$

and similarly $\mathrm{Im}(L_i) = \mathrm{Im}\big(W_V^{(i)\prime}\big)$. Thus the column spaces of $W_V^{(i)}$ and $W_V^{(i)\prime}$ coincide. Since both have full column rank d_v (A4), there exists a unique $C_i \in \mathrm{GL}(d_v)$ with

$$W_V^{(i)\prime} \;=\; W_V^{(i)} C_i.$$

Substituting back into $L_i = W_V^{(i)\prime} W_{O,i}'$ yields $W_{O,i}' = C_i^{-1} W_{O,i}$.

Query–key sector (column-space argument; A4 only). Identifiability also gives $Q_i'(K_i')^\top = Q_i K_i^\top$ for all X, i.e.

$$M_i \;:=\; W_Q^{(i)} (W_K^{(i)})^\top \;=\; W_Q^{(i)\prime} (W_K^{(i)\prime})^\top.$$

By A4, $W_K^{(i)}$ has full column rank d_k, hence $(W_K^{(i)})^\top : \mathbb{R}^{d_{\text{model}}} \to \mathbb{R}^{d_k}$ is *surjective*. Thus

$$\mathrm{Im}(M_i) \;=\; \mathrm{Im}\big(W_Q^{(i)} (W_K^{(i)})^\top\big) \;=\; \mathrm{Im}\big(W_Q^{(i)}\big),$$

and likewise $\mathrm{Im}(M_i) = \mathrm{Im}\big(W_Q^{(i)\prime}\big)$. Therefore the column spaces of $W_Q^{(i)}$ and $W_Q^{(i)\prime}$ coincide. Since both have full column rank d_k (A4), there exists a unique $A_i \in \mathrm{GL}(d_k)$ with

$$W_Q^{(i)\prime} \;=\; W_Q^{(i)} A_i.$$

Substituting back into $M_i = W_Q^{(i)\prime} (W_K^{(i)\prime})^\top$ then gives

$$W_K^{(i)\prime} \;=\; W_K^{(i)} (A_i^{-1})^\top.$$

Combining the two sectors and restoring the permutation σ proves that every gauge transformation factors as claimed. ∎

B.5. Proof of Block-Diagonality (No Cross-Head Mixing)

Lemma 23 (Transformations cannot mix heads except by permutation) *Let $P \in \mathrm{GL}(hd_v)$ act on concatenated values as $[V_1 \cdots V_h] \mapsto [V_1 \cdots V_h]P$ with compensating $W_O \mapsto P^{-1} W_O$. If this preserves* MHA *for all inputs, then P is a block permutation times a block-diagonal matrix:*

$$P \;=\; \Pi_\sigma \,\mathrm{diag}(C_1, \ldots, C_h) \quad \textit{with} \quad \sigma \in S_h,\; C_i \in \mathrm{GL}(d_v).$$

Proof Partition P into $d_v \times d_v$ blocks P_{ji}. The transformed value block for head i is

$$V_i' \;=\; \sum_{j=1}^{h} V_j\, P_{ji} \quad \text{and} \quad W_{O,i}' \;=\; \sum_{k=1}^{h} (P^{-1})_{ik}\, W_{O,k},$$

where $W_{O,i}'$ is the i-th block row of $P^{-1}W_O$ and $W_{O,k}$ is the k-th block row of W_O.

By Lemma 19, attention weights are preserved up to a permutation. Reindex heads by that permutation (fix it to id near $t = 0$ if needed), so the head-wise mixture decomposition is aligned. Then by Theorem 22, for each head i there exists $C_i \in \mathrm{GL}(d_v)$ such that the value–output pair transforms *within* head i:

$$\sum_{j=1}^{h} W_V^{(j)} P_{ji} \;=\; W_V^{(i)}\, C_i, \qquad \sum_{k=1}^{h} (P^{-1})_{ik}\, W_{O,k} \;=\; C_i^{-1}\, W_{O,i}. \tag{B.11}$$

Introduce the block-row selector matrices $E_k \in \mathbb{R}^{d_v \times (hd_v)}$ so that $W_{O,k} = E_k W_O$ and write the second equality in (B.11) as

$$\Big(\sum_{k=1}^{h} (P^{-1})_{ik} E_k\Big) W_O \;=\; C_i^{-1} E_i W_O.$$

By Assumption A8, W_O has full row rank and hence admits a right inverse W_O^+ with $W_O W_O^+ = I_{hd_v}$. Right-multiplying by W_O^+ yields

$$\sum_{k=1}^{h} (P^{-1})_{ik} E_k \;=\; C_i^{-1} E_i. \tag{B.12}$$

Since the E_k have disjoint support and are linearly independent as matrices, (B.12) forces

$$(P^{-1})_{ik} = 0 \;\; (k \neq i), \qquad (P^{-1})_{ii} = C_i^{-1}.$$

Thus P^{-1} is block-diagonal with diagonal blocks C_i^{-1}, i.e. $P = \mathrm{diag}(C_1, \ldots, C_h)$. Undoing the initial head reindexing reintroduces a (block) permutation Π_σ, completing the claim. ■

Remark 24 (Dimensional regimes) *The proof uses only A8 (full row rank of W_O) to cancel W_O via a right inverse; no head isolation (A6) is required. In the canonical case (A2: $d_{\text{model}} = hd_v$), W_O is square and invertible, so $W_O^+ = W_O^{-1}$.*

B.6. Completeness of the Gauge Group Characterization

Theorem 25 (No Additional Symmetries Exist) *The gauge group $G_{\max} = ((GL(d_k))^h \times (GL(d_v))^h) \rtimes S_h$ contains all parameter symmetries of the multi-head attention mechanism. No additional continuous or discrete symmetries exist beyond those identified in this group.*

Proof We prove completeness by showing that any purported additional symmetry would violate one of our three established constraints.

Suppose there exists a parameter transformation $\phi : \Theta_0 \to \Theta_0$ preserving the multi-head attention function that is not in $G_{\max}$. Then ϕ must satisfy:

$$\mathrm{MHA}(X; \phi(\theta)) = \mathrm{MHA}(X; \theta) \quad \forall X \in \mathbb{R}^{n \times d_{\text{model}}}, \theta \in \Theta_0 \tag{B.13}$$

By Lemma 19 (Attention Weight Identifiability), this invariance implies that attention weights are preserved up to head permutation. Therefore, ϕ must include a permutation component $\sigma \in S_h$.

After accounting for this permutation, consider the continuous component of ϕ. By Lemma 20 (Lie Algebra Characterization), any continuous one-parameter family of symmetries has its tangent vectors in $\mathfrak{g}_{\max} = \bigoplus_{i=1}^{h} \mathfrak{gl}(d_k) \oplus \bigoplus_{i=1}^{h} \mathfrak{gl}(d_v)$. This completely determines the continuous symmetries.

By Theorem 22 (Factorization), any gauge transformation must factor into independent query-key and value-output transformations. This forces:

$$\phi|_{\text{query-key}} : (W_Q^{(i)}, W_K^{(i)}) \mapsto (W_Q^{(i)} A_i, W_K^{(i)} (A_i^{-1})^\top) \tag{B.14}$$

$$\phi|_{\text{value-output}} : (W_V^{(i)}, W_{O,i}) \mapsto (W_V^{(i)} C_i, C_i^{-1} W_{O,i}) \tag{B.15}$$

for some $(A_i, C_i) \in \mathrm{GL}(d_k) \times \mathrm{GL}(d_v)$.

By Lemma 23 (Block-Diagonality), transformations cannot mix parameters across heads except through permutations. This eliminates any additional discrete symmetries beyond S_h.

Therefore, any symmetry ϕ must have the form:

$$\phi = ((A_1, \ldots, A_h), (C_1, \ldots, C_h), \sigma) \in ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h = G_{\max} \tag{B.16}$$

This contradicts our assumption that $\phi \notin G_{\max}$. Therefore, no additional symmetries exist. ■

Corollary 26 (Gauge Group is Maximal) *Among all groups acting on the parameter space that preserve the multi-head attention function, $G_{\max}$ is the unique maximal group on the generic stratum Θ_0.*

Proof Any group H of parameter symmetries satisfies $H \subseteq G(\theta)$ for all $\theta \in \Theta_0$. By Theorem 25, $G(\theta) = G_{\max}$. Therefore $H \subseteq G_{\max}$, establishing maximality. ■

Appendix C. Architectural Extensions

C.1. LayerNorm Obstruction to Inter-Layer Coupling

Lemma 27 (LayerNorm Obstruction) *Let $\mathrm{LN}_{\gamma,\beta}(x) = \gamma \odot \widehat{x} + \beta$ with*

$$\widehat{x} := \frac{Px}{\sigma(x)}, \qquad P := I - \tfrac{1}{d}\mathbf{1}\mathbf{1}^\top, \qquad \sigma(x) := \|Px\|/\sqrt{d}.$$

Assume $\gamma \in \mathbb{R}^d$ has pairwise distinct, nonzero entries and $\beta \in \mathbb{R}^d$ is generic in the sense that (i) $\mu(\beta) := \frac{1}{d}\mathbf{1}^\top \beta \neq 0$ and (ii) each coordinate of $P\beta$ is nonzero.[1] *If $M \in \mathrm{GL}(d)$ satisfies*

$$\mathrm{LN}_{\gamma,\beta}(Mx) = M\,\mathrm{LN}_{\gamma,\beta}(x) \quad \textit{for all } x \textit{ in a nonempty open set,}$$

then $M = I$.

Proof Write $\Gamma := \mathrm{diag}(\gamma)$ and $U := \{u \in \mathbb{R}^d : \mathbf{1}^\top u = 0\}$.

Step 0 (bias fixed). For any constant vector $x = c\mathbf{1}$ we have $Px = 0$ and by definition $\mathrm{LN}(x) = \beta$. Equivariance on such inputs yields

$$\mathrm{LN}(Mc\mathbf{1}) = M\beta = \beta,$$

hence

$$M\beta = \beta. \tag{C.1}$$

Step 1 (centering invariance $\Rightarrow M\mathbf{1} \in \mathrm{span}\{\mathbf{1}\}$ and $PM = MP$). For any $u \in U$ and any $c \in \mathbb{R}$, $\mathrm{LN}(u + c\mathbf{1}) = \mathrm{LN}(u)$. By equivariance,

$$\mathrm{LN}(M(u + c\mathbf{1})) = \mathrm{LN}(Mu + c\,M\mathbf{1}) = M\,\mathrm{LN}(u),$$

so the left-hand side is independent of c. Since $\mathrm{LN}(z)$ depends only on Pz, we must have $P(Mu + c\,M\mathbf{1})$ independent of c, hence $PM\mathbf{1} = 0$ and therefore

$$M\mathbf{1} = \lambda\,\mathbf{1} \quad \text{for some } \lambda \neq 0, \qquad \text{and} \qquad PM = MP. \tag{C.2}$$

Step 2 (equivariance equation on U). For $u \in U$ we have $\mathrm{LN}(u) = \gamma \odot (u/\sigma(u)) + \beta$ and, using $PM = MP$,

$$\mathrm{LN}(Mu) = \gamma \odot \frac{PMu}{\sigma(Mu)} + \beta = \gamma \odot \frac{Mu}{\sigma(Mu)} + \beta.$$

Equivariance gives

$$\Gamma\,\frac{Mu}{\sigma(Mu)} \;=\; M\,\Gamma\,\frac{u}{\sigma(u)} \qquad (\forall u \in U \setminus \{0\}). \tag{C.3}$$

Rearranging,

$$M\Gamma u \;=\; \left(\tfrac{\sigma(u)}{\sigma(Mu)}\right)\Gamma M u. \tag{C.4}$$

The left-hand side is linear in u; the right-hand side is linear in u multiplied by the scalar $r(u) := \sigma(u)/\sigma(Mu)$. For (C.4) to hold on a nonempty open cone in U, $r(u)$ must be *constant* there (analyticity), hence on all of U:

$$\frac{\sigma(Mu)}{\sigma(u)} \equiv c \quad \text{(constant)}.$$

Therefore M is conformal on U: $\|Mu\| = c\,\|u\|$ for all $u \in U$.

1. These conditions are generic (Zariski-open); drawing (γ, β) at random satisfies them almost surely.

Step 3 (orthogonality on U and commutation with Γ). With $\sigma(Mu) = c\,\sigma(u)$, (C.3) becomes

$$\Gamma\, Mu \;=\; c\, M\, \Gamma u \qquad (\forall u \in U). \tag{C.5}$$

Polarizing the identity $\|M(u \pm v)\|^2 = c^2\|u \pm v\|^2$ shows $M|_U$ preserves inner products up to the factor c^2, hence $M|_U = c\,Q$ for some $Q \in O(U)$. Plugging this into (C.5) gives $\Gamma Q = Q\Gamma$ on U. Since γ has pairwise distinct entries, the commutant of Γ in $\mathrm{GL}(d)$ is the diagonal algebra in the standard basis; because U is Γ-invariant and $Q \in O(U)$, this forces

$$Q = \mathrm{diag}(\varepsilon_1, \ldots, \varepsilon_d)\big|_U \quad \text{with } \varepsilon_j \in \{\pm 1\}.$$

Now (C.5) with $i = j$ implies $\gamma_i \varepsilon_i = c\,\varepsilon_i \gamma_i$, hence $c = 1$. Thus $M|_U = Q$ is orthogonal and diagonal with signs.

Step 4 (bias excludes sign flips on U). Decompose $\beta = \mu(\beta)\mathbf{1} + P\beta$. Using (C.1) and (C.2),

$$M\beta = \lambda\,\mu(\beta)\,\mathbf{1} + M(P\beta) \;=\; \mu(\beta)\mathbf{1} + P\beta.$$

Projecting onto U gives $M(P\beta) = P\beta$. Since $M|_U = \mathrm{diag}(\varepsilon_j)|_U$ and $P\beta$ has no zero coordinates, each ε_j must equal $+1$. Hence

$$M|_U = I|_U. \tag{C.6}$$

Step 5 (constant direction fixed). Projecting $M\beta = \beta$ onto $\mathrm{span}\{\mathbf{1}\}$ yields $\lambda\,\mu(\beta) = \mu(\beta)$, so $\lambda = 1$ because $\mu(\beta) \neq 0$. Together with (C.6) this gives $M = I$. ∎

Remark 28 (Parameter tying and non-generic cases) *If parameters are tied across layers, or if the genericity conditions on (γ, β) fail (e.g., $\mu(\beta) = 0$ or some coordinate of $P\beta$ vanishes), the centralizer can enlarge along these measure-zero loci. Our multi-layer direct-product result therefore asserts generic equality and explicitly excludes tied/degenerate configurations.*

C.2. LayerNorm Compatibility Analysis

Proof [Proof of Corollary 6] We show that the gauge group $G_{\max}$ remains a symmetry of the complete Transformer block including LayerNorm by analyzing both Pre-LN and Post-LN configurations.

Post-LN Configuration: The block computes:

$$Y = \mathrm{LN}(X + \mathrm{MHA}(X)) \tag{C.7}$$

Under any gauge transformation $g \in G_{\max}$:

$$\mathrm{MHA}(X; g(\theta)) = \mathrm{MHA}(X; \theta) \tag{C.8}$$

by Theorem 2. Therefore, the input to LayerNorm is:

$$X + \mathrm{MHA}(X; g(\theta)) = X + \mathrm{MHA}(X; \theta) \tag{C.9}$$

Since LayerNorm is a deterministic function of its input:

$$\mathrm{LN}(z) = \gamma \odot \frac{z - \mu(z)\mathbf{1}}{\sigma(z)} + \beta \tag{C.10}$$

where $\mu(z)$ and $\sigma(z)$ are the mean and standard deviation computed over the d_{model} dimension, the block output remains invariant.

Pre-LN Configuration: The block computes:

$$Y = X + \mathrm{MHA}(\mathrm{LN}(X)) \tag{C.11}$$

Let $Z = \mathrm{LN}(X)$. The gauge transformation acts only on MHA parameters, not on the normalized input. Since Theorem 2 establishes invariance for any input:

$$\mathrm{MHA}(Z; g(\theta)) = \mathrm{MHA}(Z; \theta) \tag{C.12}$$

The residual connection adds the unchanged X, preserving the total output. ■

Remark 29 (Architectural Constraint) *If LayerNorm operated on the concatenated head outputs before projection, the value-output transformations C_i would change the statistics of individual heads, breaking gauge invariance. The standard architectural choice to normalize after W_O is therefore essential, not merely conventional.*

C.3. Feed-Forward Network Gauge Structure

Lemma 30 (FFN gauge group is generically $S_{d_{ff}}$) *For a one-hidden-layer FFN block with a non-polynomial activation (e.g., GELU) and generic weights (i.e., outside a measure-zero exceptional set), the only gauge transformations of the hidden layer that preserve the block's function are hidden-unit permutations. In particular, the FFN gauge group is $G_{\text{FFN}} = S_{d_{ff}}$ on the generic stratum.*

Proof The FFN computes

$$\mathrm{FFN}(Z) = \mathrm{GELU}(ZW_1 + b_1)W_2 + b_2. \tag{C.13}$$

Consider a hidden-layer reparameterization $M \in \mathrm{GL}(d_{ff})$ acting as

$$W_1 \mapsto W_1 M, \qquad b_1 \mapsto b_1 M, \qquad W_2 \mapsto M^{-1} W_2. \tag{C.14}$$

For functional invariance under such a gauge transformation we require

$$\mathrm{GELU}(HM)M^{-1} = \mathrm{GELU}(H) \tag{C.15}$$

for all pre-activations $H = ZW_1 + b_1$.

Using the scalar form

$$\mathrm{GELU}(x) = x\,\Phi(x) = \frac{x}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right], \tag{C.16}$$

we see that GELU is non-polynomial and not homogeneous: for generic x and $\lambda \neq 1$, $\mathrm{GELU}(\lambda x) \neq \lambda\,\mathrm{GELU}(x)$. In particular, the invariance condition already rules out any nontrivial global scaling $M = \lambda I$ with $\lambda \neq 1$.

On the other hand, for permutation matrices $P \in S_{d_{ff}}$, GELU commutes element-wise:

$$\mathrm{GELU}(HP) = \mathrm{GELU}(H)P. \tag{C.17}$$

Under the corresponding reparameterization $W_1 \mapsto W_1 P$, $b_1 \mapsto b_1 P$, $W_2 \mapsto P^\top W_2$, the output becomes

$$\big(\mathrm{GELU}(H)P\big)(P^\top W_2) + b_2 = \mathrm{GELU}(H)W_2 + b_2, \tag{C.18}$$

so permutations preserve the function.

Moreover, for non-polynomial activations and generic weights, the hidden units produce linearly independent real-analytic features of the input. Classical identifiability results for single-hidden-layer networks then imply that any functional equivalence between two such networks arises only up to hidden-unit permutations (and, in some formulations, trivial rescalings that can be absorbed into W_2); see, for example, [13]. Since the GELU-based invariance condition $\mathrm{GELU}(HM)M^{-1} = \mathrm{GELU}(H)$ already rules out nontrivial scalings, the only remaining admissible reparameterizations of the form

$$W_1 \mapsto W_1 M, \qquad b_1 \mapsto b_1 M, \qquad W_2 \mapsto M^{-1} W_2$$

that preserve the FFN function for all inputs are permutation matrices $M = P \in S_{d_{ff}}$. Hence $G_{\mathrm{FFN}} = S_{d_{ff}}$ on the generic stratum. ■

C.4. Multi-Layer Direct Product Structure

Theorem 31 (No inter-layer gauge coupling) *In an L-layer Transformer, gauge transformations cannot couple parameters across different layers.*

Proof Consider a potential coupling where layer l's output is scaled by $M \in \mathrm{GL}(d_{\mathrm{model}})$ to affect layer $l+1$.

Residual incompatibility: Layer $l+1$ receives:

$$H_{\mathrm{input}}^{(l+1)} = H^{(l)} + \mathrm{Block}^{(l)}(H^{(l)}) \tag{C.19}$$

If we transform to produce $M \cdot \mathrm{Block}^{(l)}(H^{(l)})$, the total becomes:

$$H^{(l)} + M \cdot \mathrm{Block}^{(l)}(H^{(l)}) \neq M \cdot (H^{(l)} + \mathrm{Block}^{(l)}(H^{(l)})) \tag{C.20}$$

unless $M = I$ or acts on both terms, which is impossible since $H^{(l)}$ comes from the previous layer.

LayerNorm non-equivariance: LayerNorm computes:

$$\mathrm{LN}(x) = \gamma \odot \frac{x - \mu(x)}{\sigma(x)} + \beta \tag{C.21}$$

For $M \neq I$:

$$\mathrm{LN}(Mx) \neq M \cdot \mathrm{LN}(x) \tag{C.22}$$

because M changes the statistics non-uniformly.

Concrete demonstration: Consider $M = \mathrm{diag}(2, 1, \ldots, 1)$ and $x_1 = [1, 0, \ldots, 0]^\top$ with $d_{\text{model}} = 768$. Then $Mx_1 = [2, 0, \ldots, 0]^\top$ with:

$$\mu(Mx_1) = 2/768 \approx 0.0026 \tag{C.23}$$

$$\sigma(Mx_1) = \sqrt{(768-1) \cdot 4/768^2} \approx 0.0721 \tag{C.24}$$

The first component of $\mathrm{LN}(Mx_1)$ equals $(2 - 2/768)/0.0721 \approx 27.71$.

For $\mathrm{LN}(x_1)$:

$$\mu(x_1) = 1/768 \approx 0.0013 \tag{C.25}$$

$$\sigma(x_1) = \sqrt{767/768^2} \approx 0.0361 \tag{C.26}$$

The first component equals $(1 - 1/768)/0.0361 \approx 27.67$.

Thus $M \cdot \mathrm{LN}(x_1)$ has first component $2 \times 27.67 = 55.34$, while $\mathrm{LN}(Mx_1)$ has first component 27.71. Since $27.71 \neq 55.34$, LayerNorm blocks inter-layer gauge propagation. ■

Appendix D. RoPE Commutant and Reduced Gauge Group

Proposition 32 (RoPE commutant on each 2D plane) *Let* $R(\theta) = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right) \in \mathrm{SO}(2)$ *and let* $\mathcal{H}_j = \{R(\omega_j p) : p \in \mathbb{Z}\} \subset \mathrm{SO}(2)$ *act on the* j*-th* 2*D rotational plane of the RoPE map. If* $\omega_j/\pi \notin \mathbb{Q}$ *(the standard RoPE case), then the commutant in* $\mathrm{GL}(2, \mathbb{R})$ *is*

$$\mathrm{Comm}(\mathcal{H}_j) = \{\, aI_2 + bJ \;:\; a, b \in \mathbb{R},\; J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right) \,\} \;\cong\; \mathrm{GL}(1, \mathbb{C}).$$

Proof Since ω_j/π is irrational, the subgroup $\mathcal{H}_j$ is dense in $\mathrm{SO}(2)$. Thus commuting with $\mathcal{H}_j$ is equivalent to commuting with all of $\mathrm{SO}(2)$. Complexify the real 2D rotation representation via the isomorphism $\mathbb{R}^2 \cong \mathbb{C}$, where $R(\theta)$ acts as multiplication by $e^{i\theta}$. By Schur's lemma (irreducibility over $\mathbb{C}$), the commutant is the full scalar algebra $\mathbb{C}$; viewed over $\mathbb{R}$ this is precisely $\{aI_2 + bJ\}$. ■

Theorem 33 (RoPE gauge group reduction) *With* $d_k/2$ *independent* 2*D planes, the query–key commutant is*

$$\mathcal{C}_{\text{RoPE}} \cong \prod_{j=1}^{d_k/2} \mathrm{GL}(1, \mathbb{C}),$$

hence has real dimension d_k*. Consequently,*

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

Remark 34 (Intuition) *RoPE couples the query and key coordinates across positions by applying position-dependent rotations on each* 2*D plane. The only linear maps that remain as gauge symmetries are therefore those that commute with all such rotations, which is exactly the commutant* $\mathcal{C}_{\text{RoPE}}$ *described above; this is what reduces the query–key gauge freedom from* $\mathrm{GL}(d_k)$ *to a* d_k*-dimensional subgroup.*

Remark 35 (Frequency collisions) *If* $r > 1$ *planes share identically the same frequency schedule (non-generic), the isotypic component has multiplicity* r *and the commutant enlarges to* $\mathrm{GL}(r, \mathbb{C})$ *on that block. Standard RoPE uses distinct frequencies, yielding* $r = 1$ *generically.*

Appendix E. Optimization Implications

E.1. Complete Hessian Nullspace Analysis

Proposition 36 (Hessian nullspace structure) *At any critical point $\theta^* \in \Theta_0$, the Hessian $\nabla^2 L(\theta^*)$ has nullspace dimension at least $h(d_k^2 + d_v^2)$ per layer, with null directions corresponding to $\mathfrak{g}_{\max}$.*

Proof Since gauge transformations preserve the network function, they preserve the loss:

$$L(g(\theta)) = L(\theta) \quad \forall g \in G_{\max}, \theta \in \Theta_0 \tag{E.1}$$

Consider a one-parameter subgroup $g_t = \exp(tX)$ where $X \in \mathfrak{g}_{\max}$. The loss remains constant along this curve:

$$L(g_t(\theta^*)) = L(\theta^*) \quad \forall t \in \mathbb{R} \tag{E.2}$$

Taking derivatives:

$$\frac{d}{dt}\Big|_{t=0} L(g_t(\theta^*)) = \nabla L(\theta^*) \cdot v_X = 0 \tag{E.3}$$

Taking the second derivative:

$$\frac{d^2}{dt^2}\Big|_{t=0} L(g_t(\theta^*)) = v_X^\top \nabla^2 L(\theta^*) v_X = 0 \tag{E.4}$$

Remark. The Hessian equality above uses that $\nabla L(\theta^*) = 0$. At noncritical points θ, invariance still gives $\nabla L(\theta) \cdot v_X = 0$ for all $X \in \mathfrak{g}_{\max}$, but the second derivative along a gauge orbit satisfies $\frac{d^2}{dt^2}\big|_{t=0} L(g_t(\theta)) = v_X^\top \nabla^2 L(\theta)\, v_X + \nabla L(\theta) \cdot \theta''(0)$, so the Hessian need not vanish along v_X unless $\nabla L(\theta) = 0$.

The generators of $\mathfrak{g}_{\max}$ are:

$$X_{pq}^{(i,k)}: \quad \delta W_Q^{(i)} = W_Q^{(i)} E_{pq}, \quad \delta W_K^{(i)} = -W_K^{(i)} E_{pq}^\top \tag{E.5}$$

$$Y_{rs}^{(i,v)}: \quad \delta W_V^{(i)} = W_V^{(i)} E_{rs}, \quad \delta W_{O,i} = -E_{rs} W_{O,i} \tag{E.6}$$

These directions are linearly independent by the free action of $G_{\max}$ on Θ_0. Hence they span a null subspace of dimension $h(d_k^2 + d_v^2)$ per layer. ■

Remark 37 (Quotient Non-degeneracy) *If the Hessian restricted to the gauge-orthogonal (horizontal) subspace is non-degenerate at θ^* (i.e., the quotient Hessian on $\Theta/G_{\max}$ is non-singular), then the nullspace dimension equals exactly $h(d_k^2 + d_v^2)$ per layer.*

E.2. Gradient Orthogonality and Quotient Space Dynamics

Proposition 38 (Gradient orthogonality to gauge orbits)

The gradient $\nabla L(\theta)$ is orthogonal to gauge orbits: for any $X \in \mathfrak{g}_{\max}$,

$$\langle \nabla L(\theta), v_X \rangle = 0 \tag{E.7}$$

Proof The invariance $L(g_t(\theta)) = L(\theta)$ holds for all t. Differentiating at $t = 0$ yields the orthogonality condition. ■

Theorem 39 (Optimization in quotient space)
Gradient descent on the parameter space Θ is equivalent to optimization on the quotient space $\Theta/G_{\max}$ of gauge-inequivalent configurations.

Proof Since gradients are orthogonal to gauge orbits, the gradient flow equation:

$$\frac{d\theta}{dt} = -\nabla L(\theta) \tag{E.8}$$

preserves the gauge orbit. The flow factors through the quotient map $\pi : \Theta \to \Theta/G_{\max}$, inducing a flow on the quotient space with effective dimension:

$$\dim_{\text{eff}} = \dim(\Theta) - \dim(G_{\max}) = \dim(\Theta) - L \cdot h(d_k^2 + d_v^2) \tag{E.9}$$

■

Remark 40 (Formal quotient viewpoint) *Theorem 39 is stated in the standard formal sense used in the optimization and geometry literature: passing from Θ to the quotient $\Theta_0/G_{\max}$ means identifying parameters that differ only by a gauge transformation and considering gradient flow on the corresponding equivalence classes. A fully rigorous treatment would require endowing $\Theta_0/G_{\max}$ with a manifold structure and working with Riemannian gradients on this quotient; we do not pursue these technical details here, since they do not affect the practical implications.*

Theorem 41 (Mode Connectivity via the Identity Component) *Let $G_{\max}^0$ denote the identity component of $G_{\max}$. If $\theta_2 = g(\theta_1)$ with $g \in G_{\max}^0$ and $\theta_1, \theta_2 \in \Theta_0$, then there exists a continuous path $\gamma : [0,1] \to \Theta_0$ with $\gamma(0) = \theta_1$, $\gamma(1) = \theta_2$, and $L(\gamma(t)) = L(\theta_1)$ for all $t \in [0,1]$.*

Proof For $\theta_2 = g(\theta_1)$ with $g \in G_{\max}^0$, we construct the path through the one-parameter subgroup connecting the identity to g.

Since $G_{\max} = ((\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h) \rtimes S_h$, we first handle the continuous part. Any element $(A_1, \ldots, A_h, C_1, \ldots, C_h) \in (\mathrm{GL}(d_k))^h \times (\mathrm{GL}(d_v))^h$ in the identity component can be written as:

$$(A_i, C_i) = (\exp(X_i), \exp(Y_i)) \tag{E.10}$$

for some $(X_i, Y_i) \in \mathfrak{gl}(d_k) \times \mathfrak{gl}(d_v)$.

Define the path:

$$\gamma(t) = g_t(\theta_1) \text{ where } g_t = ((\exp(tX_1), \ldots, \exp(tX_h)), (\exp(tY_1), \ldots, \exp(tY_h)), \mathrm{id}) \tag{E.11}$$

By the proof of Corollary 21, these flows preserve the multi-head attention function:

$$\mathrm{MHA}(X; \gamma(t)) = \mathrm{MHA}(X; \theta_1) \quad \forall t \in [0,1] \tag{E.12}$$

Therefore:

$$L(\gamma(t)) = L(\theta_1) \quad \forall t \in [0,1] \tag{E.13}$$

For elements outside the identity component (e.g., nontrivial permutations or reflections), see the remark below. ■

Remark. Elements outside $G_{\max}^0$ (e.g., nontrivial permutations or reflections) lie in disconnected components of $G_{\max}$ and therefore do not admit a continuous path of gauge transformations from the identity. Theorem 41 thus applies only to gauge transformations in the identity component.

E.3. Gauge-Aware Optimization

Definition 42 (Gauge-Fixed Optimization) *A gauge-fixing condition $\Psi : \Theta \to \mathbb{R}^{\dim(G_{\max})}$ with $\Psi(\theta) = 0$ selects a unique representative from each gauge orbit. Constrained optimization on the gauge slice $\{\theta : \Psi(\theta) = 0\}$ eliminates redundant parameters.*

Definition 43 (Gauge-Invariant Gradient) *The gauge-invariant gradient projects the full gradient onto the orthogonal complement of gauge orbits:*

$$\nabla_{inv} L = \nabla L - \Pi_{\mathfrak{g}} \nabla L \tag{E.14}$$

where $\Pi_{\mathfrak{g}}$ denotes projection onto the Lie algebra $\mathfrak{g}_{\max}$ at the current parameters.

Proposition 44 (Natural Gradient and Gauge Structure) *The natural gradient, using the Fisher information metric, automatically accounts for gauge structure by inducing a Riemannian metric on the quotient space $\Theta/G_{\max}$. The Fisher metric is degenerate along gauge directions with zero eigenvalues corresponding to $\mathfrak{g}_{\max}$.*

Proof The Fisher information matrix $F_{ij} = \mathbb{E}[\partial_i \log p(y|x,\theta) \partial_j \log p(y|x,\theta)]$ depends only on the conditional distribution $p(y|x,\theta)$, which is invariant under gauge transformations. Therefore, $F v_X = 0$ for any gauge direction v_X. ■

E.4. Canonical Forms and Model Merging

Proposition 45 (Deterministic Canonical Form; Partial Slice) *On the generic stratum Θ_0 (Assumptions A1–A4), there exists a deterministic gauge-fixing that sets, for each head i:*

1. $(W_V^{(i)})^\top W_V^{(i)} = I_{d_v}$ *(*orthonormal value columns — reduces $\mathrm{GL}(d_v)$ to $O(d_v)$*);*
2. $(W_Q^{(i)})^\top W_Q^{(i)} = I_{d_k}$ *(*orthonormal query columns — reduces $\mathrm{GL}(d_k)$ to $O(d_k)$*);*
3. *Heads are ordered by $\|W_K^{(i)}\|_F$ in descending order (*breaks S_h*).*

This gauge-fixing reduces the value–output sector from $\mathrm{GL}(d_v)$ to $O(d_v)$ and the query–key sector from $\mathrm{GL}(d_k)$ to $O(d_k)$ per head, removing $\frac{d_v(d_v+1)}{2} + \frac{d_k(d_k+1)}{2}$ continuous degrees of freedom per head while leaving a residual $\frac{d_v(d_v-1)}{2} + \frac{d_k(d_k-1)}{2}$ orthogonal freedom.

Proof *Query–key sector.* Since $W_Q^{(i)}$ has full column rank (A4), compute the thin QR factorization with positive diagonal

$$W_Q^{(i)} = Q_Q^{(i)} R_Q^{(i)},$$

where $Q_Q^{(i)} \in \mathbb{R}^{d_{\text{model}} \times d_k}$ has orthonormal columns and $R_Q^{(i)} \in \mathrm{GL}(d_k)$ is upper triangular with $\mathrm{diag}(R_Q^{(i)}) > 0$. Choose $A_i := (R_Q^{(i)})^{-1}$ and define

$$\widetilde{W}_Q^{(i)} := W_Q^{(i)} A_i = Q_Q^{(i)}, \qquad \widetilde{W}_K^{(i)} := W_K^{(i)} (A_i^{-1})^\top = W_K^{(i)} (R_Q^{(i)})^\top.$$

Then $(\widetilde{W}_Q^{(i)})^\top \widetilde{W}_Q^{(i)} = I_{d_k}$ and the bilinear form $Q_i K_i^\top = X W_Q^{(i)} (W_K^{(i)})^\top X^\top$ is preserved, since we have applied exactly the gauge transformation $(W_Q^{(i)}, W_K^{(i)}) \mapsto (W_Q^{(i)} A_i, W_K^{(i)} (A_i^{-1})^\top)$. The positive-diagonal QR convention fixes the sign ambiguity of $Q_Q^{(i)}$, but a residual $O(d_k)$ freedom remains: for any orthogonal $A' \in O(d_k)$, $(Q_Q^{(i)} A')^\top (Q_Q^{(i)} A') = I_{d_k}$, so the orthonormal-columns condition does not uniquely determine the gauge.

Value–output sector. Analogously, compute $W_V^{(i)} = Q_V^{(i)} R_V^{(i)}$ (thin QR with $\mathrm{diag}(R_V^{(i)}) > 0$) and choose $C_i := (R_V^{(i)})^{-1}$:

$$\widetilde{W}_V^{(i)} := W_V^{(i)} C_i = Q_V^{(i)}, \qquad \widetilde{W}_{O,i} := C_i^{-1} W_{O,i} = R_V^{(i)} W_{O,i}.$$

Then $(\widetilde{W}_V^{(i)})^\top \widetilde{W}_V^{(i)} = I_{d_v}$ and $V_i W_{O,i}$ is preserved, since this corresponds to the standard value–output gauge transformation $(W_V^{(i)}, W_{O,i}) \mapsto (W_V^{(i)} C_i, C_i^{-1} W_{O,i})$. As in the query–key case, any orthogonal $C' \in O(d_v)$ applied on the right of $\widetilde{W}_V^{(i)}$ preserves the orthonormal-columns condition, so a residual $O(d_v)$ freedom remains.

Ordering. In canonical form, we drop tildes and write the resulting parameters again as $W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}, W_{O,i}$. Generically the Frobenius norms $\|W_K^{(i)}\|_F$ are distinct (ties occur on a measure-zero set). Sorting heads in descending order of $\|W_K^{(i)}\|_F$ therefore yields a unique permutation that fixes the S_h freedom. If an exact tie occurs, we break it using any fixed deterministic rule (e.g., a lexicographic order on additional invariants), which preserves determinism of the canonical representative.

These steps reduce the continuous $\mathrm{GL}(d_v)$ freedom per head to $O(d_v)$ and the $\mathrm{GL}(d_k)$ freedom to $O(d_k)$; the discrete S_h freedom is fixed generically by head ordering. ■

Conclusion. The orthonormal-columns conditions reduce each $\mathrm{GL}(d_v)$ and $\mathrm{GL}(d_k)$ to residual orthogonal groups $O(d_v)$ and $O(d_k)$, respectively, but cannot eliminate these orthogonal freedoms entirely. Geometrically, requiring $W^\top W = I$ constrains only the *symmetric* part of the Lie algebra $\mathfrak{gl}(d)$ (dimension $\frac{d(d+1)}{2}$), leaving the *antisymmetric* part $\mathfrak{so}(d)$ (dimension $\frac{d(d-1)}{2}$) unconstrained. An alternative gauge-fixing via Gram balancing (setting $(W_Q^{(i)})^\top W_Q^{(i)} = (W_K^{(i)})^\top W_K^{(i)}$) achieves the same reduction from $\mathrm{GL}(d_k)$ to $O(d_k)$. Hence the canonical form is unique up to head permutations and the residual $O(d_k) \times O(d_v)$ per head (plus tie cases when norms coincide).

Remark 46 (Residual discrete symmetries) *On the measure-zero locus where two or more heads share identical $\|W_K^{(i)}\|_F$, residual permutations among those tied heads remain; a fixed tie-break removes this ambiguity deterministically.*

Proposition 47 (Gauge-Aligned Averaging) *For models $\theta_1, \theta_2 \in \Theta_0$ with identical architecture, meaningful averaging requires gauge alignment:*

$$\theta_{\text{avg}} = \frac{\theta_1 + g^*(\theta_2)}{2}, \qquad g^* \in \arg\min_{g \in G_{\max}} \|\theta_1 - g(\theta_2)\|_F.$$

Proof Different trainings typically land at gauge-equivalent representatives of (near-)identical functions. The minimization over $G_{\max}$ decomposes into continuous (A_i, C_i) parts and a discrete permutation component, by the factorization and block-diagonality results of Appendix B. In particular, the canonical slice (Proposition 45) reduces the continuous part from $\mathrm{GL}(d_k) \times \mathrm{GL}(d_v)$ to $O(d_k) \times O(d_v)$ per head, so the minimizer involves a permutation alignment between canonical representatives together with residual orthogonal adjustments within each head.

Concretely, map both models to canonical form head-wise (orthonormalize $W_Q^{(i)}$ and $W_V^{(i)}$ via thin QR with positive diagonals, and transport $W_K^{(i)}$ and $W_{O,i}$ accordingly) to obtain canonical parameters $\widetilde{\theta}_1, \widetilde{\theta}_2$. The alignment problem over heads becomes a linear assignment with cost

$$D_{ij} \;=\; \big\|\widetilde{W}_{K,1}^{(i)} - \widetilde{W}_{K,2}^{(j)}\big\|_F^2 \;+\; \big\|\widetilde{W}_{V,1}^{(i)} - \widetilde{W}_{V,2}^{(j)}\big\|_F^2 \;+\; \big\|\widetilde{W}_{O,1,i} - \widetilde{W}_{O,2,j}\big\|_F^2,$$

(or any fixed gauge-invariant metric on canonical blocks). The optimal permutation σ solving this assignment (e.g., via the Hungarian algorithm) yields the discrete part of g^*. The residual $O(d_k) \times O(d_v)$ freedom per head may be further reduced by Procrustes alignment if desired. Averaging $\theta_{\mathrm{avg}} = \frac{1}{2}\big(\widetilde{\theta}_1 + \sigma \cdot \widetilde{\theta}_2\big)$ then respects the dominant gauge structure and remains close (in Frobenius norm) to each model's orbit representative. ■

Appendix F. Experimental Validation Details

Appendix G. GPT-2: Assumption Verification, Gauge Invariance, and Canonicalization

This section validates the gauge symmetry framework on the complete GPT-2 model family using publicly available HuggingFace checkpoints. We consider GPT-2 Small (124M), GPT-2 Medium (355M), GPT-2 Large (774M), and GPT-2 XL (1.5B parameters). For each model we:

1. verify structural assumptions on the learned attention weights (Assumptions A1–A3, A4, A7, A8);
2. test gauge invariance using random valid and invalid transformations;
3. apply a canonicalization procedure that fixes the V/O gauge and balances the Q/K sector; and
4. check end-to-end exactness of the canonicalized model at the logit and generation levels.

All computations use IEEE single-precision arithmetic (FP32, $\varepsilon_{\mathrm{mach}}^{\mathrm{FP32}} = 2^{-23} \approx 1.19 \times 10^{-7}$) on NVIDIA H100 GPUs. Unless otherwise stated, reported errors are relative Frobenius norms of the form

$$\frac{\|Y' - Y\|_F}{\|Y\|_F + \varepsilon_{\mathrm{mach}}^{\mathrm{FP32}}},$$

where Y and Y' are the original and transformed outputs. Gauge transformations are always applied to *pretrained* weights; no random or synthetic models are used.

G.1. Experimental Setup

Pretrained models. We load GPT-2 checkpoints from HuggingFace via the `GPT2LMHeadModel` interface and extract the attention projections W_Q, W_K, W_V, W_O and associated biases on a per-head basis. The relevant architectural hyperparameters are summarized in Table 2. We retain the native FP32 weights and do not perform any finetuning or re-initialization.

Model	Params	Layers L	Heads h	d_{model}	d_k	d_v
GPT-2 Small	124M	12	12	768	64	64
GPT-2 Medium	355M	24	16	1024	64	64
GPT-2 Large	774M	36	20	1280	64	64
GPT-2 XL	1.5B	48	25	1600	64	64

Table 2: GPT-2 configurations used for empirical validation. All models share $d_k = d_v = 64$ and a standard context length of 1024 tokens.

Assumption verification. For every layer ℓ and head i we compute singular values and condition numbers of the query, key, value, and output projections using `torch.linalg.svdvals`. Assumption A4 requires that $W_Q^{(\ell,i)}, W_K^{(\ell,i)}, W_V^{(\ell,i)} \in \mathbb{R}^{d_{\text{model}} \times d_k}$ have full column rank; Assumption A8 requires that $W_O^{(\ell,i)} \in \mathbb{R}^{d_v \times d_{\text{model}}}$ have full row rank. Assumption A7 concerns linear independence of the bilinear forms $M_i^{(\ell)} = W_Q^{(\ell,i)} (W_K^{(\ell,i)})^\top$ across heads.

Gauge invariance tests. To probe gauge invariance we sample random, well-conditioned transformations $A \in \mathrm{GL}(d_k)$ and $C \in \mathrm{GL}(d_v)$ for each test. The implementation constructs A and C as small perturbations of random orthogonal matrices, ensuring that A and C are invertible with bounded condition numbers. For each model we perform 20 tests; each test selects one (ℓ, i) pair (cycling over layers and heads) and a random input vector $x \sim \mathcal{N}(0, I_{d_{\text{model}}})$, then compares

$$\text{logit} = qk^\top, \qquad \text{output} = vW_O$$

with their gauge-transformed counterparts

$$\text{logit}' = (qA)(kA^{-\top})^\top, \qquad \text{output}' = (vC)(C^{-1}W_O),$$

where $q = xW_Q$, $k = xW_K$, and $v = xW_V$. We record relative errors for both the scalar attention logit and the value-output product.

Invalid transformations. We also evaluate three families of deliberately invalid transformations:

1. *Asymmetric Q/K transform*: distinct matrices applied to Q and K, breaking the $A^{-\top}$ compatibility condition.

2. *Wrong inverse*: using the same matrix A for both Q and K instead of $(A^{-1})^\top$, again violating the invariance condition on $qk^\top$.

3. *V/O mismatch*: transforming V by a nontrivial C but leaving W_O unchanged, so that vW_O is not preserved.

For each model we apply a single instance of each invalid transform and measure the relative error in the attention logit.

Canonicalization and exactness. Finally, we apply a canonicalization procedure that (i) orthonormalizes the value projections W_V so that $W_V^\top W_V \approx I_{d_v}$ for each head, and (ii) balances the Q/K sector by whitening the associated covariance matrices. This procedure corresponds to a particular choice of gauge in the maximal symmetry group and is implemented via SPD matrix square roots and their inverses (see Appendix E.3.1 for algorithmic details). We then compare the original and canonicalized models on ten natural-language prompts using:

- teacher-forced logit comparisons at every position, and
- greedy decoding (temperature 0) to check that generated tokens match exactly.

G.2. Assumption Verification on Pretrained Weights

Projection matrices (A4, A8). Table 3 summarizes the singular-value ranges and mean condition numbers for the projection matrices across all layers and heads. For each entry we report the range of the minimum singular value $\sigma_{\min}$ and the mean condition number κ.

Model	Matrix	$\sigma_{\min}$ range	mean κ
GPT-2 Small	W_Q	[0.36, 5.79]	4.00
GPT-2 Small	W_K	[0.44, 5.33]	4.36
GPT-2 Small	W_V	[0.41, 3.44]	2.88
GPT-2 Small	W_O	[0.28, 3.31]	4.87
GPT-2 Medium	W_Q	[0.37, 4.33]	5.43
GPT-2 Medium	W_K	[0.40, 4.41]	5.08
GPT-2 Medium	W_V	[0.57, 3.64]	2.64
GPT-2 Medium	W_O	[0.12, 3.67]	4.38
GPT-2 Large	W_Q	[0.46, 2.38]	3.12
GPT-2 Large	W_K	[0.45, 2.48]	3.01
GPT-2 Large	W_V	[0.28, 1.58]	2.29
GPT-2 Large	W_O	[0.31, 2.04]	2.58
GPT-2 XL	W_Q	[0.50, 2.00]	3.28
GPT-2 XL	W_K	[0.49, 2.23]	3.08
GPT-2 XL	W_V	[0.26, 1.59]	2.04
GPT-2 XL	W_O	[0.31, 1.75]	2.52

Table 3: Singular-value ranges and mean condition numbers for projection matrices in pretrained GPT-2 models. All matrices have full rank, with $\sigma_{\min} \geq 0.12$ in every case and mean condition numbers between 2.0 and 5.5, indicating that the learned projections are well-conditioned for gauge transformations.

Across all four models every instance of W_Q, W_K, W_V has full column rank and every instance of W_O has full row rank. The smallest singular values are substantially bounded away from zero, and no rank-deficient projection matrices are observed. Assumptions A4 and A8 therefore hold uniformly on the pretrained GPT-2 family.

Bilinear-form independence (A7). For each layer ℓ we form the bilinear-form matrices $M_i^{(\ell)} = W_Q^{(\ell,i)}(W_K^{(\ell,i)})^\top \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ and stack them into a matrix in $\mathbb{R}^{h \times d_{\text{model}}^2}$. We then compute its rank, minimum singular value, and condition number. The results are summarized in Table 4.

Model	all layers full rank?	full-rank layers	$\min_\ell \sigma_{\min}$	$\max_\ell \kappa$
GPT-2 Small	Yes	12/12	26.04	13.13
GPT-2 Medium	No	22/24	23.93	15.92
GPT-2 Large	No	30/36	7.51	9.37
GPT-2 XL	No	32/48	7.70	8.28

Table 4: Bilinear-form independence (Assumption A7) for pretrained GPT-2 models. "Full-rank layers" counts how many layers have stacked bilinear forms of rank exactly h.

For GPT-2 Small all layers satisfy Assumption A7 exactly: the h bilinear forms are linearly independent in every layer. For the larger models we observe a small number of layers (2 out of 24 for GPT-2 Medium, 6 out of 36 for GPT-2 Large, and 16 out of 48 for GPT-2 XL) where the numerical rank of the stacked bilinear forms falls slightly below h. However, even in these layers the smallest nonzero singular values remain large ($\sigma_{\min} \geq 7.5$) and condition numbers are modest ($\kappa \leq 16$). We interpret these cases as benign, training-induced near-degeneracies (e.g., two heads learning almost identical bilinear forms) rather than structural violations of the generic Assumption A7. Empirically, they have no adverse effect on gauge invariance or canonicalization, as shown below.

G.3. Gauge Invariance on Pretrained GPT-2

Per-model summary. Table 5 reports mean and maximum relative errors for attention outputs and logits under random valid gauge transformations, computed over the 20 tests per model.

Model	mean output error	max output error	mean logit error	max logit error
GPT-2 Small	1.26×10^{-6}	1.63×10^{-6}	1.78×10^{-6}	1.23×10^{-5}
GPT-2 Medium	1.21×10^{-6}	2.16×10^{-6}	2.03×10^{-6}	2.11×10^{-5}
GPT-2 Large	1.20×10^{-6}	1.94×10^{-6}	2.05×10^{-6}	1.14×10^{-5}
GPT-2 XL	1.24×10^{-6}	2.50×10^{-6}	2.38×10^{-6}	2.04×10^{-5}

Table 5: Relative errors under *valid* gauge transformations on pretrained GPT-2 models (20 tests per model). Errors are relative Frobenius norms for outputs and scalar relative errors for logits.

Across all models, the mean relative attention-output error lies in the range 1.2–1.3×10^{-6}, while the maximum error is at most 2.5×10^{-6}. For logits the mean relative error is on the order of 2×10^{-6} and the maximum error is at most 2.1×10^{-5}. In terms of FP32 machine epsilon $\varepsilon_{\text{mach}}^{\text{FP32}}$, these correspond to roughly 10–21 $\varepsilon_{\text{mach}}^{\text{FP32}}$ for attention outputs and up to about $180\,\varepsilon_{\text{mach}}^{\text{FP32}}$ for logits, well within the expected accumulation band for deep matrix computations in single precision. We did not observe any numerical instabilities or outliers beyond these ranges.

Distributional analysis. To characterize the error distribution, we aggregate the 20 tests per model (80 tests total) and compute empirical percentiles. The results appear in Table 6.

Quantity	min	median	90th	95th	99th	max
Attention outputs	7.47×10^{-7}	1.22×10^{-6}	1.53×10^{-6}	1.63×10^{-6}	2.23×10^{-6}	2.50×10^{-6}
Logits	0	7.57×10^{-7}	4.31×10^{-6}	9.11×10^{-6}	2.06×10^{-5}	2.11×10^{-5}

Table 6: Distribution of relative errors under valid gauge transformations, aggregated over all four GPT-2 models (80 tests total).

For attention outputs, the median relative error is $1.22 \times 10^{-6} \approx 10\,\varepsilon_{\text{mach}}^{\text{FP32}}$, the 95th percentile is $1.63 \times 10^{-6} \approx 14\,\varepsilon_{\text{mach}}^{\text{FP32}}$, and the maximum is $2.50 \times 10^{-6} \approx 21\,\varepsilon_{\text{mach}}^{\text{FP32}}$. For logits, the median relative error is $7.6 \times 10^{-7} \approx 6\,\varepsilon_{\text{mach}}^{\text{FP32}}$, the 95th percentile is $9.1 \times 10^{-6} \approx 76\,\varepsilon_{\text{mach}}^{\text{FP32}}$, and the maximum is $2.1 \times 10^{-5} \approx 177\,\varepsilon_{\text{mach}}^{\text{FP32}}$. Errors are tightly clustered with no heavy tails: 99% of tests lie below 2.3×10^{-6} for attention outputs and below 2.1×10^{-5} for logits.

Invalid transformations. The invalid transformations produce errors that are large in absolute terms and orders of magnitude above the valid-gauge regime. Table 7 summarizes the relative errors in the attention logit for each invalid transform.

Model	Asymmetric Q/K	Wrong inverse	V/O mismatch
GPT-2 Small	4011.7	2824.9	120.5
GPT-2 Medium	2935.2	1018.5	31.7
GPT-2 Large	24.9	2319.2	26.3
GPT-2 XL	398.2	353.9	4.0

Table 7: Relative errors in the attention logit under three families of *invalid* transformations. All errors are $O(10^0)$–$O(10^3)$, in stark contrast to the 10^{-6}–10^{-5} errors observed for valid transforms.

Even the mildest invalid transformation (V/O mismatch on GPT-2 XL) produces a relative error of ≈ 4.0, while asymmetric or mis-matched Q/K transforms often produce errors in the thousands. The dramatic separation between the valid-gauge error band and the invalid-transform errors provides strong empirical evidence that the gauge group identified in the theory is indeed maximal: perturbations outside this group almost always produce $O(1)$–$O(10^3)$ changes in the attention logits.

G.4. Canonicalization Effects

We now quantify the effect of the canonicalization procedure on the learned attention weights. For each model we track:

- the Frobenius norm $\|W_V^\top W_V - I_{d_v}\|_F$ (value orthonormality);
- the Q/K imbalance $\|S_Q - S_K\|_F / \|S_Q\|_F$, where S_Q and S_K are empirical covariances of Q and K;
- the condition numbers of S_Q and S_K.

Table 8 reports layer- and head-averaged metrics before and after canonicalization.

	$\|W_V^\top W_V - I\|_F$		Q/K imbalance		$\kappa(S_Q)$		$\kappa(S_K)$	
Model	before	after	before	after	before	after	before	after
GPT-2 Small	82.8	1.51×10^{-6}	0.507	1.25×10^{-2}	25.6	20.9	39.8	20.4
GPT-2 Medium	76.5	1.46×10^{-6}	0.453	1.69×10^{-2}	87.1	65.0	96.3	66.4
GPT-2 Large	8.8	1.46×10^{-6}	0.563	8.80×10^{-3}	11.5	6.9	10.3	6.8
GPT-2 XL	8.2	1.47×10^{-6}	0.546	9.45×10^{-3}	14.0	9.6	14.2	9.4

Table 8: Canonicalization metrics averaged over layers and heads in pretrained GPT-2 models.

On all four models the value projections are far from orthonormal in the pretrained checkpoints: the mean value of $\|W_V^\top W_V - I\|_F$ ranges from ≈ 8 to ≈ 83, with per-layer values reaching as high as 216 on GPT-2 Small and 166 on GPT-2 Medium. Canonicalization reduces this quantity to $\approx 1.5 \times 10^{-6}$ across all models, consistent with numerical roundoff in FP32.

The Q/K imbalance metric exhibits similar improvement. Pretrained models have mean imbalance in the range 0.45–0.56, with some layers exceeding 3.2 (GPT-2 Small) or 9.9 (GPT-2 Large). After canonicalization the mean imbalance falls into the 0.009–0.017 range, a 30–60$\times$ reduction.

Finally, canonicalization systematically reduces the condition numbers of the Q/K covariance matrices. For GPT-2 Small and Medium the mean condition numbers decrease from $(\kappa(S_Q), \kappa(S_K)) \approx (26, 40)$ and $(87, 96)$ to $(21, 20)$ and $(65, 66)$, respectively. For GPT-2 Large and XL they decrease from roughly $(12, 10)$ and $(14, 14)$ to $(6.9, 6.8)$ and $(9.6, 9.4)$. In all cases the conditioned covariances remain well within numerically stable regimes while exhibiting substantially improved symmetry.

G.5. Exactness of the Canonicalized Model

To confirm that canonicalization is a *pure gauge choice* that leaves the input–output behavior unchanged, we compare the original and canonicalized models on ten diverse natural-language prompts. Table 9 summarizes the logit-level and generation-level results.

Model	mean logit difference	max logit difference	greedy decode match
GPT-2 Small	1.50×10^{-5}	1.34×10^{-4}	100%
GPT-2 Medium	2.32×10^{-5}	1.91×10^{-4}	100%
GPT-2 Large	1.18×10^{-6}	4.20×10^{-5}	100%
GPT-2 XL	1.20×10^{-6}	3.05×10^{-5}	100%

Table 9: Exactness verification between original and canonicalized GPT-2 models. Differences are measured in absolute logits under teacher forcing; generation match reports the fraction of prompts for which greedy decoding produces identical token sequences.

For all four GPT-2 models we obtain a 100% greedy-decoding match rate: canonicalized models produce *exactly* the same token sequences as the original checkpoints on all ten prompts. At the logit level the overall maximum difference ranges from 3.1×10^{-5} (GPT-2 XL) to 1.9×10^{-4} (GPT-2 Medium), corresponding to at most $\approx 1.6 \times 10^3\, \varepsilon_{\text{mach}}^{\text{FP32}}$. Average logit differences are much smaller, between 1.2×10^{-6} and 2.3×10^{-5}.

These results demonstrate that the canonicalization procedure is functionally exact up to FP32 numerical precision: it leaves the sequence distribution unchanged while substantially improving the conditioning and symmetry properties of the learned attention parameters.

G.6. Summary

The GPT-2 experiments provide a comprehensive, real-world validation of the gauge symmetry framework:

- Pretrained GPT-2 weights satisfy the structural assumptions (A4, A7, A8) with significant numerical margin, apart from a handful of benign, near-degenerate layers in the largest models.
- Valid gauge transformations preserve attention outputs and logits to within $O(10)$–$O(200)$ FP32 machine epsilons, with tightly concentrated error distributions and no outliers.
- Invalid transformations produce relative errors in the range $O(1)$–$O(10^3)$, empirically reinforcing the maximality of the gauge group characterized in the theory.
- Canonicalization dramatically improves value orthonormality, Q/K balance, and covariance conditioning while preserving both logits and generated sequences up to FP32 numerical precision.

Together, these findings show that the theoretical gauge structure is not only mathematically exact but also faithfully realized in production-scale transformer language models with hundreds of millions to billions of parameters.

Appendix H. Practical Implementation Details

H.1. Gauge-Fixing Algorithms

To reduce gauge redundancy, we fix parameters to a canonical form using numerically stable per-head QR factorizations. This construction reduces the continuous $\mathrm{GL}(d_k)$ and $\mathrm{GL}(d_v)$ freedoms to residual orthogonal groups $O(d_k) \times O(d_v)$ per head and then breaks the remaining S_h permutation symmetry by a deterministic head ordering. When biases are present, we transform them covariantly (Proposition 12) to preserve exact functional invariance.

Algorithm 1 Gauge-Fixing to Canonical Form (QR-based)

Input: MHA parameters $\{W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}, W_{O,i}\}_{i=1}^h$ (optional biases $\{b_Q^{(i)}, b_K^{(i)}, b_V^{(i)}\}$ and global b_O) with full column rank (A4)
Output: Canonical parameters with continuous gauge freedom reduced to $O(d_k) \times O(d_v)$ per head

Per-head canonicalization (apply the following steps for each head $i \in \{1, \dots, h\}$):
  Query–Key canonicalization
    Compute a thin QR factorization $W_Q^{(i)} = Q_Q R_Q$.
    Enforce the positive-diagonal convention for R_Q by flipping column/row signs so $\mathrm{diag}(R_Q) > 0$.
    Set $W_Q^{(i)} \leftarrow Q_Q$ (now $(W_Q^{(i)})^\top W_Q^{(i)} = I_{d_k}$).
    Set $W_K^{(i)} \leftarrow W_K^{(i)} R_Q^\top$ (preserves $Q_i K_i^\top$ and reduces the $\mathrm{GL}(d_k)$ gauge freedom to $O(d_k)$).
    `// (If present) transform Q/K biases covariantly`
    if $b_Q^{(i)}$ exists **then** $b_Q^{(i)} \leftarrow b_Q^{(i)} R_Q^{-1}$ **end if** ▷ Prop. 12
    if $b_K^{(i)}$ exists **then** $b_K^{(i)} \leftarrow b_K^{(i)} R_Q^\top$ **end if** ▷ Prop. 12
  Value–Output canonicalization
    Compute a thin QR factorization $W_V^{(i)} = Q_V R_V$.
    Enforce the positive-diagonal convention for R_V (make $\mathrm{diag}(R_V) > 0$).
    Set $W_V^{(i)} \leftarrow Q_V$ (now $(W_V^{(i)})^\top W_V^{(i)} = I_{d_v}$).
    Set $W_{O,i} \leftarrow R_V W_{O,i}$ (preserves $V_i W_{O,i}$ and reduces the $\mathrm{GL}(d_v)$ gauge freedom to $O(d_v)$).
    `// (If present) transform V/O biases covariantly`
    if $b_V^{(i)}$ exists **then** $b_V^{(i)} \leftarrow b_V^{(i)} R_V^{-1}$ **end if** ▷ Prop. 12
    ▷ Global output bias b_O is added after projection and remains unchanged

Break permutation symmetry (generic case)
Compute norms$[i] \leftarrow \|W_K^{(i)}\|_F$ for $i = 1, \dots, h$.
Let $\sigma \leftarrow$ ArgSort(norms, `descending`).
Reorder all head-indexed parameters by σ (ties can be broken lexicographically).

Remark 48 (Positive-diagonal QR convention) *Many `QR` implementations may return R with mixed signs on the diagonal. To enforce the positive-diagonal convention (which yields a unique factorization), set $s_j = \mathrm{sign}(R[j,j])$ (take $s_j = 1$ if $R[j,j] = 0$), then update $Q[:,j] \leftarrow s_j\, Q[:,j]$ and $R[j,:] \leftarrow s_j\, R[j,:]$. Apply this to (Q_Q, R_Q) and (Q_V, R_V).*

Remark 49 (Why this is a deterministic partial slice) *With $W_Q^{(i)}$ and $W_V^{(i)}$ reduced to column-orthonormal forms via a fixed positive-diagonal QR convention, any further gauge transformation $(A_i, C_i) \in \mathrm{GL}(d_k) \times \mathrm{GL}(d_v)$ that preserves $(W_Q^{(i)})^\top W_Q^{(i)} = I_{d_k}$ and $(W_V^{(i)})^\top W_V^{(i)} = I_{d_v}$ must satisfy $A_i \in O(d_k)$ and $C_i \in O(d_v)$. Thus the continuous gauge group is reduced from $\mathrm{GL}(d_k) \times \mathrm{GL}(d_v)$ to $O(d_k) \times O(d_v)$ per head, exactly as in Proposition 45.*

Ordering by $\|W_K^{(i)}\|_F$ removes the residual S_h ambiguity; ties occur on a measure-zero set and are broken lexicographically.

Remark 50 (Complexity) *Each thin QR costs $\mathcal{O}(d_{\text{model}} d_k^2)$ or $\mathcal{O}(d_{\text{model}} d_v^2)$. The per-layer cost is $\mathcal{O}\big(h(d_{\text{model}} d_k^2 + d_{\text{model}} d_v^2)\big)$ and is typically negligible relative to a forward pass.*

H.2. Gauge-Aware Optimization

Projecting gradients onto directions orthogonal to the gauge orbits eliminates updates that do not change the model's function. Conceptually, this is motivated by Proposition 38 (gradients are orthogonal to gauge orbits) and Theorem 39 (optimization proceeds on the quotient). The routine below implements a practical approximate projection used in our experiments.

Algorithm 2 Approximate Gauge-Orthogonal Gradient Projection

Input: Gradient $\nabla L = \{\nabla_{W_Q^{(i)}} L, \nabla_{W_K^{(i)}} L, \nabla_{W_V^{(i)}} L, \nabla_{W_{O,i}} L\}_{i=1}^h$

Output: Projected gradient $\nabla_\perp L$

Per head $i = 1, \dots, h$**:**

Query–Key projection

$X_i \leftarrow (W_Q^{(i)})^\dagger \nabla_{W_Q^{(i)}} L$

$Y_i \leftarrow -(W_K^{(i)})^\dagger \nabla_{W_K^{(i)}} L$

$Z_i \leftarrow \frac{1}{2}(X_i + Y_i^\top)$ // *symmetrize*

$\nabla^\perp_{W_Q^{(i)}} L \leftarrow \nabla_{W_Q^{(i)}} L - W_Q^{(i)} Z_i$

$\nabla^\perp_{W_K^{(i)}} L \leftarrow \nabla_{W_K^{(i)}} L + W_K^{(i)} Z_i^\top$

Value–Output projection

$U_i \leftarrow (W_V^{(i)})^\dagger \nabla_{W_V^{(i)}} L$

$V_i \leftarrow -\nabla_{W_{O,i}} L\, (W_{O,i})^\dagger$

$T_i \leftarrow \frac{1}{2}(U_i + V_i)$

$\nabla^\perp_{W_V^{(i)}} L \leftarrow \nabla_{W_V^{(i)}} L - W_V^{(i)} T_i$

$\nabla^\perp_{W_{O,i}} L \leftarrow \nabla_{W_{O,i}} L + T_i W_{O,i}$

Remark 51 (Exact Euclidean projection via Sylvester equations) *The exact Frobenius-orthogonal projection onto the gauge-orthogonal subspace solves, per head, a Sylvester equation. For the query–key sector, the optimal Z_i minimizing $\|\nabla_{W_Q^{(i)}} L - W_Q^{(i)} Z_i\|_F^2 + \|\nabla_{W_K^{(i)}} L + W_K^{(i)} Z_i^\top\|_F^2$ satisfies*

$$G_Q^{(i)} Z_i + Z_i G_K^{(i)} = (W_Q^{(i)})^\top \nabla_{W_Q^{(i)}} L - (\nabla_{W_K^{(i)}} L)^\top W_K^{(i)},$$

where $G_Q^{(i)} := (W_Q^{(i)})^\top W_Q^{(i)}$ and $G_K^{(i)} := (W_K^{(i)})^\top W_K^{(i)}$. An analogous equation holds for the value–output sector with $G_V^{(i)} := (W_V^{(i)})^\top W_V^{(i)}$ and $G_O^{(i)} := W_{O,i} W_{O,i}^\top$. The symmetrization

heuristic in Algorithm 2 avoids the $\mathcal{O}(d_k^3 + d_v^3)$ Sylvester solves per head at the cost of a small approximation error.

H.3. Model Merging via Gauge Alignment

Independent trainings land at different points on the same gauge orbit. Aligning gauges before averaging yields meaningful interpolations. The procedure below fixes both models to a canonical gauge, matches heads, and fine-tunes per-head alignment.

Algorithm 3 Gauge Alignment for Model Merging

Input: Two models θ_1, θ_2 with the same architecture
Output: Aligned model θ_2' minimizing $\|\theta_1 - \theta_2'\|_F$

Step 1: Gauge-fix both models to canonical form
$\theta_1 \leftarrow \textsc{GaugeFix}(\theta_1)$; $\theta_2 \leftarrow \textsc{GaugeFix}(\theta_2)$
Step 2: Find optimal head permutation (Hungarian algorithm)
$D_{ij} \leftarrow \|W_{Q,1}^{(i)} - W_{Q,2}^{(j)}\|_F + \|W_{K,1}^{(i)} - W_{K,2}^{(j)}\|_F$
$\sigma \leftarrow \textsc{HungarianAlgorithm}(D)$; permute heads of θ_2 by σ
Step 3: Per-head fine alignment (query–key Procrustes)
for $i = 1, \ldots, h$: $A_i \leftarrow \arg\min_A \|W_{Q,1}^{(i)} - W_{Q,2}^{(i)} A\|_F$; apply (A_i, I_{d_v}) to head i of θ_2

H.4. Computational Complexity

All flop counts below are *per layer* and for dense operations; batch size multiplies costs by B, and multi-layer models scale linearly in L. We use the canonical dimensions $d_{\text{model}} = h\, d_v$ when helpful for intuition, but the formulas hold without this identity.

Table 10: Computational complexity of gauge operations versus standard operations (per layer).

Operation	**Complexity per Layer**
Gauge transformation	$O\big(h(d_k^2\, d_{\text{model}} + d_v^2\, d_{\text{model}})\big)$
Gauge-fixing (QR-based)	$O\big(h(d_{\text{model}}\, d_k^2 + d_{\text{model}}\, d_v^2)\big)$
Gradient projection	$O\big(h(d_k^2\, d_{\text{model}} + d_v^2\, d_{\text{model}})\big)$
Model alignment	$O\big(h^3 + h(d_{\text{model}}\, d_k^2 + d_{\text{model}}\, d_v^2)\big)$
Forward pass (attention + MLP)	$O\big(n^2 d_{\text{model}} + n d_{\text{model}}^2\big)$

For typical values (n=512, d_{model}=768, h=12, d_k=d_v=64), the QR-based gauge-fixing and the gauge-orthogonal gradient projection add a small fraction of the FLOPs of a single forward pass (dominated by the $n^2 d_{\text{model}}$ attention term), and are therefore negligible in practice for common sequence lengths.