W_K, W_V IS PROBABLY ALL YOU NEED: ON THE NECESSITY OF THE QUERY, KEY, AND VALUE WEIGHT TRIPLET IN SELF-ATTENTION TRANSFORMERS

Marko Karbevski[*] **Antonij Mijoski**[†]
Institut de Recherche Mathématique Avancée (IRMA)
Université de Strasbourg

ABSTRACT

We theoretically investigate whether the Query, Key, Value weight triplet can be reduced in encoder-only and decoder-only transformers. Under mild assumptions, we prove that Query, Key or Value weights are redundant and can be replaced with the identity matrix, reducing attention parameters by 25%. If applied to the Query or Key weights, this also simplifies optimization: attention logits become linear rather than quadratic in learned weights. Validating the Query weight removal on decoder-only GPT-style small models trained from scratch, we find that with adjusted attention scaling and weight decay, reduced models match baseline performance despite fewer parameters. Training remains stable at over $3\times$ lower weight decay, suggesting Query weight elimination provides implicit regularization. Our analysis has also led us to a structural expressivity boundary: in the mathematically tractable ReLU setting, skip connections push MLPs into a generically disjoint function class at fixed width. These findings motivate investigation across modalities and at scale, where the observed stability and efficiency gains may prove most consequential.

1 INTRODUCTION

Training and deploying transformer-based language models (Vaswani et al., 2017) remains computationally expensive (Samsi et al., 2023), motivating architectural optimizations (Tay et al., 2022) including quantization (Ma et al., 2024), efficient attention (Choromanski et al., 2020; Wang et al., 2020), weight sharing (Shazeer, 2019; Ainslie et al., 2023; Lan et al., 2020), and normalization streamlining and block restructuring (He & Hofmann, 2024; Heimersheim, 2024; Baroni et al., 2025; Zhu et al., 2025). Recent work has shown that normalization layers (Heimersheim, 2024; Baroni et al., 2025) and attention parameters (He & Hofmann, 2024) can be rearranged and simplified with minimal performance impact, suggesting current architectures may be overparameterized.

We investigate redundancy within the attention mechanism itself: *is the Query-Key-Value weight triplet necessary?* The key insight is that attention depends on the input X only through the products XW_Q, XW_K, XW_V. This enables a telescoping construction where each layer's basis transformation prepares the input for the next, ultimately replacing W_Q with the identity matrix throughout the network. We focus on W_Q for its compatibility with KV caching and GQA; the theory applies identically to W_K or W_V, whose empirical evaluation we leave to future work.

Contributions. We take a theory-first approach: we prove that Query weight elimination is possible under simplifying assumptions, then validate empirically in full architectural complexity.

- **Theory (Section 4, Appendices):** We first observe that multi-head attention has intrinsic redundancy: any block-diagonal transformation can be absorbed into the Query-Key pair without changing the output (Proposition 3.2). This redundancy is parametrized by a $(h \cdot d_k^2)$-dimensional manifold (one invertible $d_k \times d_k$ matrix per head), present in the full model with no simplifying assumptions.

[*]marko.karbevski@gmail.com
[†]antonij.mijoski@unistra.fr, antonijmijo@gmail.com

Under the additional assumption of no normalization layers, we prove that W_Q can be set to identity while preserving input-output mappings for (i) any single layer in transformers with full skip connections and untied embedding/LM-head weights (Theorem 4.2, the "Single-Layer Free Lunch"), (ii) all layers in weight-shared transformers (Theorem 4.3), and (iii) all layers in transformers with skip connections around attention only (Theorem 4.4). Notably, the Single-Layer Free Lunch applies to existing pretrained models once LayerNorm is removed via the techniques of Heimersheim (2024) and Baroni et al. (2025), enabling post-training Query weight elimination.

We also derive the functional form required for basis transformations to pass through LayerNorm (Appendix C), showing that LayerNorm compatibility requires the MLP to approximate a specific nonlinear compensation. Notably, the resulting obstruction is strictly milder than the per-head modulation introduced by QK-normalization (Henry et al., 2020), which poses no empirical limitation as demonstrated by numerous SOTA models (Team, 2025; Team et al., 2025; OLMo et al., 2025).

- **Index-free notation for MHA (Appendix A):** Building on variants of the Block Hadamard Products defined in Horn et al. (1991), we provide a compact notation that more closely reflects standard implementations. Under this formulation, the motivating observation of our paper—that attention depends on X only through XW_Q, XW_K, XW_V—becomes mathematically trivial.
- **Skip connections are generically unabsorbable (Appendix E):** We prove that for ReLU MLPs at fixed width, the set of weights for which the skip connection $W_2\,\mathrm{ReLU}(W_1x)+x$ can be exactly represented as $V_2\,\mathrm{ReLU}(V_1x)$ has Lebesgue measure zero. Skip and non-skip ReLU MLPs of the same width thus represent generically disjoint function classes, clarifying that skip connections provide access to a *different* region of function space rather than strictly "more" expressivity.
- **Experiments (Section 5):** We train GPT-style models (117M to 124M parameters) from scratch with $W_Q = \mathrm{Id}$, comparing against parameter-matched baselines. With hyperparameter adjustments motivated by our theoretical analysis, the reduced 117M model matches the full 124M baseline despite 8% fewer non-embedding parameters, and outperforms parameter-matched baselines at equal size. Reallocating saved parameters to the MLP significantly outperforms the full 124M baseline counterpart. Moreover, its training remains stable at $3\times$ lower weight decay.
- **Training stability:** The reduced architecture trains stably at over $3\times$ lower weight decay. We attribute this to: (i) simplification from quadratic to linear parameter dependence in attention logits, and (ii) the identity query acting as an implicit skip connection within attention.

Scope and limitations. This work establishes sufficient conditions for the elimination of Query weight matrices (W_Q). While the theoretical derivation utilizes simplified architectures, empirical validation on GPT-style models confirms that W_Q redundancy persists in practical models. These experiments at the 117M–124M parameter scale demonstrate technical feasibility and establish the empirical baseline for systematic scaling, multi-seed validation, downstream benchmarking, and validation across modalities and architectures.

2 Related Work

Architectural simplification. Graef (2024) formally proves that in skipless transformers without normalization, both W_Q and W_O can be eliminated simultaneously and leaves the question open whether elimination extends to practical architectures. We go further in several directions: (i) we analyze transformers *with* skip connections by retaining W_O to absorb basis changes, (ii) we characterize exactly when ReLU MLPs can absorb skip connections (Appendix E), (iii) we derive the functional form required for basis transformations through LayerNorm (Appendix C), and (iv) we validate through GPT-style pretraining. He & Hofmann (2024) empirically study simplified parallel attention-MLP blocks; we focus on weight elimination within the original architecture, rather than through block restructuring. Recent work showing that LayerNorm can be removed from pretrained models (Heimersheim, 2024; Baroni et al., 2025) supports our no-normalization theoretical setting.

Notably, Ji et al. (2025) show that self-attention catastrophically fails to train without skip connections around attention, validating that our Theorem 4.4 addresses the practically essential case.

Efficient attention. Grouped-Query Attention (Ainslie et al., 2023) and Multi-Query Attention (Shazeer, 2019) reduce parameters by sharing Key-Value projections. Linear attention methods (Choromanski et al., 2020; Wang et al., 2020) reduce complexity. FlashAttention (Dao et al., 2022) optimizes memory access. Our approach is orthogonal to these optimizations: Query weight elimination applies to standard, GQA, and MQA architectures alike, enabling multiplicative savings.

Weight sharing, tying, and recursive architectures. In order to save memory Press & Wolf (2017) tie embedding and Language Modeling (LM) head weights with minimal performance degradation. This technique is further adopted in GPT-2 (Radford et al., 2019) and many subsequent decoder-only models. Our theoretical work addresses both the tied and untied regimes.

Bai et al. (2019) study weight-shared transformers as implicit fixed-point solvers. ALBERT (Lan et al., 2020) shares all parameters across layers, achieving $18\times$ memory reduction at only modest performance cost. In this case the conditions of our theory are further relaxed: we systematically treat the case where all the skip connections are present, and only rely on the approximation of the change of basis through the LayerNorm. This implies that our theoretical investigation is highly relevant for this family of models. Finally, compared to weight sharing, not only do our methods reduce the memory, but also the computational footprint.

Recursive models such as Tiny Recursion Models (TRM) (Jolicoeur-Martineau, 2025), which apply a 2-layer block repeatedly, are natural candidates for Query weight elimination under Theorem 4.3. TRM uses RMSNorm, which our analysis covers (see Appendix C). Interestingly, Jolicoeur-Martineau (2025) found that embedding-head tying degraded performance, so TRM uses untied weights, satisfying our untying requirement and making Query weight elimination directly applicable.

Theoretical foundations. Theoretical analysis of transformers typically trades architectural fidelity for mathematical tractability. Yun et al. (2019) establish universal approximation; Ildiz et al. (2024) connect self-attention to Markov dynamics; Geshkovski et al. (2024; 2023) provide mean-field analyses of attention. We do the same to obtain tractable proofs, then verify our results empirically in full architectural complexity.

Parameter-efficient fine-tuning and compression. LoRA (Hu et al., 2021) and related methods (Dettmers et al., 2024) achieve efficiency through low-rank adaptation during fine-tuning. Post-training compression via pruning (Ma et al., 2023) and quantization (Ma et al., 2024) are also orthogonal to our approach. Query weight elimination applies at the architectural level, benefiting both pre-training and inference, and can be combined with these techniques for compound efficiency gains.

3 PRELIMINARIES

We establish notation for multi-head attention and state the Reparametrization Lemma underlying our theoretical analysis.

Notation. We write $\mathrm{Mat}(p, q)$ for the set of real $p \times q$ matrices and $\mathrm{GL}(d)$ for the invertible $d \times d$ matrices. Fix n (sequence length), d_{model} (embedding dimension), and h (number of attention heads) with $h | d_{\text{model}}$. Let $d_k = d_{\text{model}}/h$ denote the dimension per head. For each head $i \in \{1, \ldots, h\}$, we define head-specific weight matrices $W_Q^i, W_K^i, W_V^i \in \mathbb{R}^{d_{\text{model}} \times d_k}$, assembled into full-layer parameters $W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ by concatenation: $W_Q = (W_Q^1 | \cdots | W_Q^h)$, and similarly for W_K, W_V. The output projection is $W_O \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$.

We let $M \in \{0, -\infty\}^{n \times n}$ denote the attention mask: $M_{ij} = 0$ if token i can attend to token j, $-\infty$ otherwise. For decoder-only (causal) attention, $M_{ij} = -\infty$ when $j > i$; for encoder-only (bidirectional) attention, $M = 0$.

The output of a single attention head i is:

$$\text{head}_i(X, W_Q, W_K, W_V) = \text{softmax}\left(\frac{(XW_Q^i)(XW_K^i)^\top}{\sqrt{d_k}} + M\right)(XW_V^i),$$

the *self-attention product* is $\mathcal{S}_c(X, W_Q, W_K, W_V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) \in \mathbb{R}^{n \times d_{\text{model}}}$, and the *multi-head attention* is

$$\text{Attn}(X, W_Q, W_K, W_V, W_O) = \mathcal{S}_c(X, W_Q, W_K, W_V) \cdot W_O. \tag{1}$$

Eliminating W_Q reduces the four attention weight matrices to three, saving 25% of attention parameters. Our results rely on $W_Q \in \text{GL}(d)$. This holds for almost every square matrix (in the Lebesgue sense) (Vershynin, 2018); near-singular W_Q implies that certain directions contribute minimally to queries, and replacing W_Q with identity restores these directions.

The Reparametrization Lemma. A key observation is that $\mathcal{S}_c$ depends on X *only through* the products XW_Q, XW_K, and XW_V:

Observation 3.1 (Factorization)**.** *There exists a function g such that $\mathcal{S}_c(X, W_Q, W_K, W_V) = g(XW_Q, XW_K, XW_V)$.*

This is verified in Appendix A, where we introduce index-free Block Hadamard Products (Horn et al., 1991) under which the observation becomes mathematically trivial. The factorization allows replacing W_Q with identity via a change of basis:

Lemma 3.1 (Reparametrization Lemma)**.** *Let $f : \text{Mat}(n, d) \times \text{Mat}(d, d)^3 \to \Omega$ satisfy $f(X, W_Q, W_K, W_V) = g(XW_Q, XW_K, XW_V)$ for some function g. Then f is invariant under $(X, W_Q, W_K, W_V) \mapsto (X\Theta, \Theta^{-1}W_Q, \Theta^{-1}W_K, \Theta^{-1}W_V)$ for any $\Theta \in \text{GL}(d)$. In particular, for any (W_Q, W_K, W_V) with $W_Q \in \text{GL}(d)$, there exist $\widetilde{W}_K, \widetilde{W}_V \in \text{Mat}(d, d)$ such that*

$$f(X, W_Q, W_K, W_V) = f(XW_Q, I_d, \widetilde{W}_K, \widetilde{W}_V) \quad \forall X \in \text{Mat}(n, d).$$

Explicitly, $\widetilde{W}_K = W_Q^{-1}W_K$ and $\widetilde{W}_V = W_Q^{-1}W_V$.

Remark 1 (Alternative reductions)**.** *By the same argument, we can eliminate W_K (setting $\Theta = W_K$) or W_V (setting $\Theta = W_V$) instead, with the remaining matrices absorbing the basis change. We focus on Query weight elimination for its compatibility with KV caching and GQA.*

Additional degrees of freedom exist in the pair (W_Q, W_K):

Proposition 3.2 (Block-diagonal invariance)**.** *Let $D = \text{diag}(D_1, \ldots, D_h) \in \text{GL}(d_{\text{model}})$ be block-diagonal with $D_i \in \text{Mat}(d_k, d_k)$. Then:*

$$\mathcal{S}_c(X, W_Q, W_K, W_V) = \mathcal{S}_c(X, W_QD, W_K(D^\top)^{-1}, W_V).$$

The proof appears in Appendix A. This reveals that any block-diagonal transformation can be absorbed without affecting the output, identifying an $(h \cdot d_k^2)$-dimensional family of redundant parameters. Similar redundancies may exist between W_V and W_O; we leave this to future work.

4 Theoretical Analysis

We present results with progressively stronger conclusions, at the cost of additional architectural assumptions. Table 1 summarizes the scope of each result. Note that for encoder-only models, embeddings and output heads are never tied, so all theorems apply without any untying requirement.

4.1 Single-Head Attention

We begin with the conceptually simplest case. In single-head attention:

$$\text{Attn}(X, W_Q, W_K, W_V, W_O) = \text{softmax}\left(\tfrac{1}{c}XW_QW_K^\top X^\top\right)XW_VW_O$$

Table 1: Summary of theoretical results for W_Q elimination (i.e., setting $W_Q = \mathrm{Id}$). ✓ = allowed/present, × = must be absent/simplified. Norm: normalization layers. Skips: skip connections. MH: multi-head attention. Layers: Standard (Std) or Shared (ALBERT-style). E ↔ H: embedding/LM-head tying.

Result	Norm	Skips	MH	Layers	E ↔ H	W_Q Eliminated
Prop. 4.1	✓	✓	×	Std	Either	All layers
Thm. 4.2	×	✓	✓	Std	Untied	Any single layer
Thm. 4.3	×	✓	✓	Shared	Untied	All layers
Thm. 4.4	×	Attn only	✓	Std	Either	All layers
Experiments	✓	✓	✓	Std	Tied	All layers

Proposition 4.1 (Single-head redundancy)**.** *For single-head attention, the four weight matrices (W_Q, W_K, W_V, W_O) can be reduced to two:*

$$\mathrm{Attn}(X, W_Q, W_K, W_V, W_O) = \mathrm{Attn}(X, \mathrm{Id}, W_K W_Q^\top, W_V W_O, \mathrm{Id}).$$

This holds regardless of normalization or skip connections because the attention mechanism depends only on $W_Q W_K^\top$ and $W_V W_O$. However, weight decay during training penalizes $\|W_Q\|^2 + \|W_K\|^2 + \|W_V\|^2 + \|W_O\|^2$ rather than $\|W_Q W_K^\top\|^2 + \|W_V W_O\|^2$, potentially preventing discovery of this reduced form.

4.2 Single-Layer Query Weight Elimination

Before addressing full Query weight elimination across all layers, we show that *any single layer's* Query weights can always be eliminated in transformers without normalization, even with all skip connections present.

Theorem 4.2 (Single-Layer Free Lunch)**.** *Consider an L-layer encoder-only or decoder-only transformer without normalization, with all skip connections, and with untied weights. Let E, E_P, W_{head} be embedding, positional embedding, and output head weights, and let $(W^i_{\mathrm{MLP}_u}, W^i_{\mathrm{MLP}_d}, W^i_Q, W^i_K, W^i_V, W^i_O)_{i=1}^L$ be the layer weights.*

For any $j \in \{1, \ldots, L\}$ with $W^j_Q \in \mathrm{GL}(d)$, there exist modified weights $(\tilde{E}, \tilde{E}_P, \tilde{W}_{head}, \tilde{W}^i_{\mathrm{MLP}_u}, \tilde{W}^i_{\mathrm{MLP}_d}, \tilde{W}^i_Q, \tilde{W}^i_K, \tilde{W}^i_V, \tilde{W}^i_O)_{i=1}^L$ such that $\tilde{W}^j_Q = \mathrm{Id}$ and the transformer with weights $(\tilde{E}, \tilde{E}_P, \tilde{W}_{head}, (\tilde{W}^i_{\mathrm{MLP}_u}, \tilde{W}^i_{\mathrm{MLP}_d}, \tilde{W}^i_Q, \tilde{W}^i_K, \tilde{W}^i_V, \tilde{W}^i_O)_{i=1}^L)$ produces identical outputs to the original.

Remark 2 (Post-training application)**.** *Heimersheim (2024) and Baroni et al. (2025) show that LayerNorm can be removed from pretrained transformers (up to GPT-2 XL) with minimal degradation. For such models, our theorem applies directly post-training: a single layer's W_Q can be eliminated through weight reparametrization alone, saving 25% of that layer's attention parameters.*

The proof (Appendix F) propagates a single basis transformation $\Theta = W^j_Q$ uniformly through the entire network.

Remark 3 (Why not all layers?)**.** *To eliminate all Query weights simultaneously, layer i would need its own basis $\Theta_i = W^i_Q$, requiring blocks to satisfy $\widetilde{B}_i(X\Theta_i) = B_i(X)\Theta_{i+1}$ with potentially $\Theta_i \neq \Theta_{i+1}$. However, the MLP skip connection forces same-basis intertwining: $\widetilde{\mathrm{MLP}}_i(Z\Theta) + Z\Theta = (\mathrm{MLP}_i(Z) + Z)\Theta$. This admits two solutions: (i) share weights so $\Theta_i = \Theta$ for all i (Theorem 4.3), or (ii) remove the MLP skip, allowing $W^i_{\mathrm{MLP}_d}$ to transform between bases (Theorem 4.4).*

When layers share weights, the single-layer elimination propagates to all attention layers, as described next.

4.3 WEIGHT-SHARED TRANSFORMERS

Theorem 4.3 (Weight-Shared Query Weight Elimination)**.** *Consider an L-layer encoder-only or decoder-only transformer without normalization where all layers share the same block parameters. Let E, E_P, W_{head} be embedding, positional embedding, and output head weights, and let $(W_{\mathrm{MLP}_u}, W_{\mathrm{MLP}_d}, W_Q, W_K, W_V, W_O)$ be the shared layer weights with $W_Q \in \mathrm{GL}(d)$.*

Then there exist modified weights $(\tilde{E}, \tilde{E}_P, \tilde{W}_{head}, \tilde{W}_{\mathrm{MLP}_u}, \tilde{W}_{\mathrm{MLP}_d}, \tilde{W}_K, \tilde{W}_V, \tilde{W}_O)$ such that the transformer with weights $(\tilde{E}, \tilde{E}_P, \tilde{W}_{head}, \tilde{W}_{\mathrm{MLP}_u}, \tilde{W}_{\mathrm{MLP}_d}, \mathrm{Id}, \tilde{W}_K, \tilde{W}_V, \tilde{W}_O)$ produces identical outputs to the original.

Proof sketch. With an appropriate change of variables, each block satisfies $B = \Theta^{-1} \circ \widetilde{B} \circ \Theta$ where $\widetilde{B}$ is a reduced block, i.e. one where $W_Q = \mathrm{Id}$. Since all layers share weights, one can use the telescoping technique to obtain $B^{\circ L} = \Theta^{-1} \circ \widetilde{B}^{\circ L} \circ \Theta$. The embedding and LM head absorb the boundary transformations. □

Remark 4 (Embedding-head untying)**.** *With tied embedding/LM-head weights ($W_{LM} = E^\top$), the reduced model must* untie *these weights to maintain equivalence, since $\tilde{W}_{LM} = W_Q^{-1} E^\top \neq \tilde{E}^\top$ unless W_Q is orthogonal.*

4.4 MULTI-HEAD ATTENTION WITH RESTRICTED SKIP CONNECTIONS

Eliminating *all* layers' Query weights with standard (non-shared) parameters requires restricting the skip connection structure. We consider skip connections around attention only, setting aside MLP skip connections and LayerNorm (see Appendix C for LayerNorm analysis).

Optimization versus trainability. Skip connections confer two distinct benefits that are often conflated. *Optimization* refers to training efficiency: skip connections smooth the loss landscape (Li et al., 2018), eliminate singularities that slow convergence (Orhan & Pitkow, 2018), and enable gradient flow through very deep networks (He et al., 2016). Networks *can* train without skip connections, but more slowly and with worse final performance. He et al. (2023) show that "vanilla" transformers (without skip connections or normalization) can be trained using special modifications to self-attention, but require approximately $5\times$ more iterations to reach equivalent performance.

Trainability, by contrast, refers to whether training succeeds at all. For self-attention specifically, Ji et al. (2025) show that training *catastrophically fails* without skip connections around attention, while removing them in other components (MLPs, convolutions) is much less consequential. Our Theorem 4.4 addresses this practically essential case: transformers with skip connections around attention.

Theorem 4.4 (Attention-Skip-Only Query Weight Elimination)**.** *Consider an L-layer encoder-only or decoder-only transformer without normalization and with skip connections only around attention. Let E, E_P, W_{head} be embedding, positional embedding, and output head weights (LM head for decoder-only, task head for encoder-only), and let $(W^i_{\mathrm{MLP}_u}, W^i_{\mathrm{MLP}_d}, W^i_Q, W^i_K, W^i_V, W^i_O)_{i=1}^L$ be the layer weights.*

Then there exist modified weights $(\tilde{E}, \tilde{E}_P, \tilde{W}_{head}, \tilde{W}^i_{\mathrm{MLP}_u}, \tilde{W}^i_{\mathrm{MLP}_d}, \tilde{W}^i_K, \tilde{W}^i_V, \tilde{W}^i_O)_{i=1}^L$ such that the transformer with weights $(\tilde{E}, \tilde{E}_P, \tilde{W}_{head}, (\tilde{W}^i_{\mathrm{MLP}_u}, \tilde{W}^i_{\mathrm{MLP}_d}, \mathrm{Id}, \tilde{W}^i_K, \tilde{W}^i_V, \tilde{W}^i_O)_{i=1}^L)$ produces identical outputs to the original.

The proof (Appendix F) proceeds by induction. Setting $\Theta_i = W^i_Q$ for each layer $i \in \{1, \ldots, L\}$ and $\Theta_{L+1} = I_d$, the transformed weights are:

$$\widetilde{E} = E\Theta_1, \qquad \widetilde{E}_P = E_P\Theta_1,$$

$$\widetilde{W}^i_K = \Theta_i^{-1} W^i_K, \qquad \widetilde{W}^i_V = \Theta_i^{-1} W^i_V, \qquad \widetilde{W}^i_O = W^i_O \Theta_i,$$

$$\widetilde{W}^i_{\mathrm{MLP}_u} = \Theta_i^{-1} W^i_{\mathrm{MLP}_u}, \qquad \widetilde{W}^i_{\mathrm{MLP}_d} = W^i_{\mathrm{MLP}_d} \Theta_{i+1}.$$

Unlike Graef (2024), who eliminates both W_Q and W_O in skipless transformers, we retain W_O and use it to adapt the output of the attention to the change of basis in the skip connection around it. This allows the construction to work with skip connections around the attention.

Table 2: Model configurations and validation loss at 100k steps.

	Baseline (smaller MLP)	Baseline (smaller d_{model})	$W_Q = \text{Id}$ (117M)	Baseline (124M)	$W_Q = \text{Id}$ (larger MLP)
Trainable W_Q	✓	✓	×	✓	×
Total params	117.30M	117.92M	117.30M	124.37M	124.37M
Non-emb params	77.88M	79.73M	77.88M	84.95M	84.95M
MLP hid. mult.	$3.5\times$	$4\times$	$4\times$	$4\times$	$4.5\times$
Attention scale	$1/\sqrt{64}$	$1/\sqrt{62}$	$1/(2\sqrt{64})$	$1/\sqrt{64}$	$1/(2\sqrt{64})$
Weight decay	0.1	0.1	2^{-5}	0.1	2^{-5}
Val loss (100k)	3.026	3.027	**3.018**	3.016	**3.004**

5 EXPERIMENTS

We validate our theoretical results by training GPT-style models from scratch with $W_Q = \text{Id}$. Preliminary experiments on MLP basis transfer approximation are in Appendix D.

5.1 PRETRAINING OF GPT-STYLE MODELS

Architecture. We use Karpathy's NanoGPT (Karpathy, 2023) implementation of GPT-2/GPT-3-small (Radford et al., 2019; Brown et al., 2020): 12 layers, 12 heads, $d_{\text{model}} = 768$, MLP hidden dimension $4d_{\text{model}} = 3072$, GELU activations (Hendrycks & Gimpel, 2016), LayerNorm (Ba et al., 2016), context length 1024, GPT-2 BPE tokenizer. Compared to GPT-2 and 3, but consistent with modern frontier models and Karpathy's implementation, we omit the bias parameters.

Training. We train for 100k gradient steps (of 600k in the full NanoGPT schedule) on OpenWebText (Gokaslan & Cohen, 2019) ($\sim$9B tokens) with mixed-precision training. We use AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, learning rate 6×10^{-4} with 2k steps of warmup and cosine decay to 6×10^{-5}, gradient clipping at 1.0, and $\sim$490k tokens per gradient step. To ensure a fair comparison, we pre-generate all training and evaluation batch indices from a fixed random seed and reuse them across all model variants, so every configuration sees identical data in identical order. Training uses a single NVIDIA RTX 5090 GPU with FlashAttention (Dao et al., 2022). All configurations use tied embedding/LM-head weights.

Model variants. To isolate the effect of Query weight elimination from parameter count differences, we compare five configurations (Table 2): two parameter-matched baselines (smaller MLP or smaller d_{model}), the standard 124M baseline, and two reduced architectures ($W_Q = \text{Id}$) with parameters either matching the 117M baselines or reallocated to a larger MLP. Following scaling law conventions (Hoffmann et al., 2022), Table 2 and Figure 1 report non-embedding parameters.

Practical adjustments. Two modifications are necessary for the reduced architecture:

1. **Attention Scaling Correction:** We adopt a scaling factor of $\frac{1}{2\sqrt{d_k}}$ instead of the standard $\frac{1}{\sqrt{d_k}}$. The intuition: with $W_Q = \text{Id}$, queries are coordinate slices of the input rather than learned projections. At initialization, this yields attention scores with approximately $1.8\times$ larger standard deviation than the baseline (see Appendix B for the full derivation). The factor of $\frac{1}{2}$ compensates for this to prevent early softmax saturation.

2. **Weight Decay:** We reduce the weight decay coefficient from 0.1 to $2^{-5} \approx 0.03$. Our theoretical results establish that the remaining parameters have sufficient capacity to jointly encode their original functions and the compensating basis transformations. In standard architectures, this capacity is latent: suppressed by weight decay, which restricts the degrees of freedom. The reduced architecture's stability at lower weight decay suggests that the remaining weights are engaging this previously regularized capacity.

Results. Table 2 and Figure 1 summarize our findings:

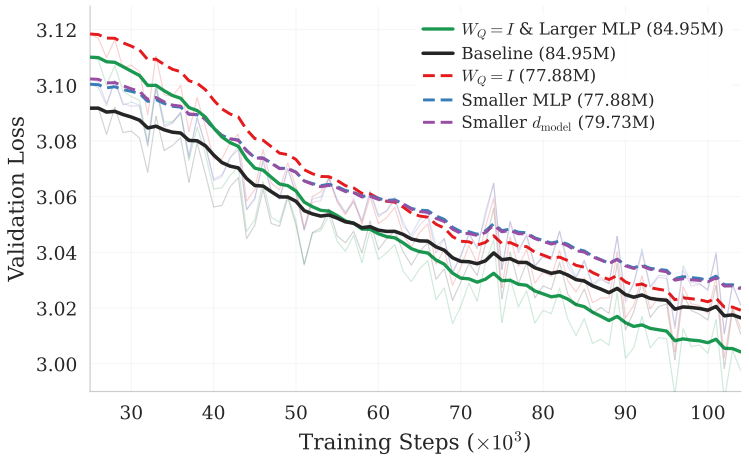

Figure 1: Validation loss during training (non-embedding parameters in parentheses).

1. **Query weights are redundant.** The reduced model ($W_Q = \text{Id}$, 117M) achieves comparable performance to the full baseline (124M) in validation loss, despite having 8% fewer non-embedding parameters.
2. **Parameter reallocation improves performance.** When saved parameters are reallocated to the MLP ($4.5 \times d_{\text{model}}$ hidden dimension), the reduced architecture achieves the *best* validation loss (3.004 vs. 3.016 for the baseline).
3. **Reduced model beats parameter-matched baselines.** At equal total parameter count (117M), the reduced model outperforms both the smaller-MLP and smaller-d_{model} baselines, suggesting Query weight elimination is preferable to naive parameter reduction.
4. **Training is stable.** Loss curves track closely throughout training with smooth convergence at over $3\times$ lower weight decay.

6 Discussion and Conclusion

Implicit regularization and complementary regimes. Setting $W_Q = \text{Id}$ provides implicit regularization: attention logits become linear rather than quadratic in learned parameters, and every coordinate slice directly participates in attention, potentially encouraging more uniform gradient flow. Weight decay protects against loss divergence in modern LLMs beyond classical regularization (D'Angelo et al., 2024); that we train stably at $3\times$ lower values suggests the reduced architecture provides inherent stability. More broadly, our theoretical and empirical results occupy complementary regimes: theory demonstrates achievability under mild assumptions in a highly generic setting; practice faces harder conditions but enjoys greater freedom, as training finds its own path to a good minimum.

Limitations. Our experiments validate the core result over 100k of the 600k training schedule at the 117M–124M parameter scale. Full training runs across multiple seeds, larger scales, and downstream benchmarks represent natural next steps.

Extensions and future directions. The reduction mechanism relies only on linear projection structure and extends to: Rotary Position Embeddings (Su et al., 2021) (fixed rotations satisfy the condition of the Reparametrization Lemma 3.1), Grouped-Query Attention (Ainslie et al., 2023) (transformations act per query-head group), and Mixture-of-Experts (elimination applies per-expert). The analysis extends without any modification to encoder-only architectures, with potentially larger relative savings. Technical directions include untied embedding configurations and QK-normalization compatibility (Henry et al., 2020).

While we target W_Q for its compatibility with KV caching and GQA, insights from parameter-efficient fine-tuning suggest the Key matrix has the smallest effect on performance (Hu et al., 2021), making W_K potentially an even better candidate for structural simplification.

Our results demonstrate that *linear* Query projections provide no benefit over identity. However, *nonlinear* Query transformations with a skip connection, i.e., $Q(X) = X + N(X; \theta)$ where N is a nonlinearity with approximately the same parameter count as W_Q, may offer additional expressivity while preserving training stability. Conversely, if KV caching is not required (as in encoder-only models), applying such nonlinearities to W_V could be equally promising. Investigating Query, Key, or Value weight elimination in Hyper-Connection architectures (Zhu et al., 2024; Xie et al., 2025) presents another interesting direction.

Conclusion. We prove that W_Q can be eliminated through basis transformations under simplified assumptions, reducing attention parameters by 25%. Models with $W_Q = \mathrm{Id}$ match baselines when appropriately tuned, and *exceed* them when saved parameters are reallocated to the MLP—consistent with our analysis identifying MLP expressivity as a limiting factor (Appendices C, E), and pointing to architectural redundancy in the Query-Key-Value triplet.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers at the ICLR DeLTA workshop for their helpful comments and suggestions, as well as Nils Graef (OpenMachine), Borjan Geshkovski (INRIA/Sorbonne), François Yvon (CNRS/Sorbonne) and Yiping Ji (University of Adelaide) for helpful discussions on this work. The first author would also like to thank Igor Ševo (HTEC/UniBL) for discussions on the attention mechanism. We also thank Dimitar Peshevski (UKIM) for suggestions on parameter analysis for hyperparameter improvements in future work. This work was initiated by the first author while at the HTEC Group (Karbevski & HTEC Group, 2024), and we thank the team for their support.

REPRODUCIBILITY

Code and checkpoints are available at `https://github.com/MarkoKarbevski/Wqkv_necessity`.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

Sanjeev Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Luca Baroni, Galvin Khara, Joachim Schaeffer, Marat Subkhankulov, and Stefan Heimersheim. Transformers don't need layernorm at inference time: Scaling layernorm removal to gpt-2 xl and the implications for mechanistic interpretability. *arXiv preprint arXiv:2507.02559*, 2025. doi: 10.48550/arXiv.2507.02559. URL `https://arxiv.org/abs/2507.02559`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Francesco D'Angelo, Maksym Andriushchenko, Aditya Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning? In A. Globerson, L. Mackey,

D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 23191–23223. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/29496c942ed6e08ecc469f4521ebfff0-Paper-Conference.pdf.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Nils Graef. Transformer tricks: Removing weights for skipless transformers. *arXiv preprint arXiv:2404.12362*, 2024. URL https://arxiv.org/abs/2404.12362.

J. Elisenda Grigsby and Kathryn Lindsey. On transversality of bent hyperplane arrangements and the topological expressiveness of ReLU neural networks. *SIAM Journal on Applied Algebra and Geometry*, 6(2):216–242, 2022.

Bobby He and Thomas Hofmann. Simplifying transformer blocks. In *International Conference on Representation Learning*, 2024.

Bobby He, James Martens, Guodong Zhang, Aleksandar Botev, Andrew Brock, Samuel L Smith, and Yee Whye Teh. Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation. In *International Conference on Learning Representations*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Stefan Heimersheim. You can remove gpt2's layernorm by fine-tuning. *arXiv preprint arXiv:2409.13710*, 2024. doi: 10.48550/arXiv.2409.13710. URL https://arxiv.org/abs/2409.13710. Presented at the ATTRIB and Interpretable AI workshops, NeurIPS 2024.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Alex Henry, Prudhvi Raj Dachapally, Shubham Paber, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35:30016–30030, 2022.

Roger A. Horn, Roy Mathias, and Yoshihiro Nakamura. Inequalities for unitarily invariant norms and bilinear matrix products. *Linear and Multilinear Algebra*, 30(4):303–314, November 1991. ISSN 1563-5139. doi: 10.1080/03081089108818114. URL http://dx.doi.org/10.1080/03081089108818114.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Muhammed Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to Markov models: Unveiling the dynamics of generative transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20955–20982. PMLR, 2024.

Yiping Ji, Hemanth Saratchandran, Peyman Moghadam, and Simon Lucey. Always skip attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Alexia Jolicoeur-Martineau. Less is more: Recursive reasoning with tiny networks. *arXiv preprint arXiv:2510.04871*, 2025.

Marko Karbevski and HTEC Group. HTEC AI Open Day Skopje — possible modifications to the attention mechanism. YouTube video, 2024. Available at: `https://www.youtube.com/watch?v=UjTHj-cFJOA`. Accessed: October 19, 2025.

Andrej Karpathy. NanoGPT. `https://github.com/karpathy/nanoGPT`, 2023.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations (ICLR)*, 2020. URL `https://arxiv.org/abs/1909.11942`.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era of 1-bit llms: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 2024.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in Neural Information Processing Systems*, 36:21702–21720, 2023.

Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2924–2932, 2014.

Team OLMo et al. Olmo 3, 2025.

A. Emin Orhan and Xaq Pitkow. Skip connections eliminate singularities. In *International Conference on Learning Representations*, 2018.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 157–163, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://aclanthology.org/E17-2025`.

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

Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–9. IEEE, 2023.

Noam M. Shazeer. Fast transformer decoding: One write-head is all you need. *ArXiv*, abs/1911.02150, 2019. URL https://api.semanticscholar.org/CorpusID:207880429.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6), 2022. doi: 10.1145/3530811. URL https://dl.acm.org/doi/10.1145/3530811.

Gemma Team et al. Gemma 3 technical report, 2025.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on Learning Theory*, pp. 1517–1539. PMLR, 2016.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Kuai Yu, Liang Zhao, Shangyan Zhou, Zhean Xu, Zhengyan Zhang, Wangding Zeng, Shengding Hu, Yuqing Wang, Jingyang Yuan, Lean Wang, and Wenfeng Liang. mHC: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*, 2025.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, and Xun Zhou. Hyper-connections. *arXiv preprint arXiv:2409.19606*, 2024.

Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, and Zhuang Liu. Transformers without normalization. *arXiv preprint arXiv:2503.10622*, 2025. doi: 10.48550/arXiv.2503.10622. URL https://arxiv.org/abs/2503.10622. CVPR 2025.

A Block-Notation for Multi-Head Attention

We proceed as follows:

First, we prove Observation 3.1 using the standard notation for the convenience of the reader.

Second, we introduce the Block Hadamard Products (as variants of those introduced in (Horn et al., 1991)) to model the multi-head mechanism. In this notation, the result becomes mathematically trivial: it is manifestly clear that the self-attention product $\mathcal{S}_c$ is a function of XW_Q, XW_K, and

XW_V alone, as the block-wise operations never access X except through these projections. The structural redundancy of the query weights is thus rendered immediate.

The notation is motivated by and provides an index-free formalization of standard implementations (e.g., (Karpathy, 2023; Vaswani et al., 2017)) which use XW_Q, XW_K, XW_V (after a split from XW_{QKV}), and reshape them into per-head blocks $(XW_Q)^1, \ldots, (XW_Q)^h$ (and similarly for XW_K, XW_V). Our notation formalizes the transition: from the full products XW_Q, XW_K, XW_V to the (parallel) per-head processing.

We remind the reader that we have fixed n (sequence length), d_{model} (embedding dimension), h (number of attention heads) and $d_k = d_{\text{model}}/h$ (dimension per head).

Slice operator. Let m, p be two integers such that $0 < m < p \leq d_{\text{model}}$, and in what follows, we will select them to be of the form $m = a \times d_k$, $p = (a+1) \times d_k$ for $0 < a < h$. We define the slice operator as the canonical projection $P_{[m,p]} : \mathbb{R}^{d_{\text{model}}} \ni x = (x_1, \ldots, x_{d_{\text{model}}}) \mapsto P_{[m,p]}(x) = xP_{[m,p]} = (x_m, x_{m+1}, \ldots, x_p)$, and we remind the reader that we consider vectors as row vectors here. For head i we write $P_i = P_{[(i-1)\times d_k+1,\ i\times d_k]}$.

Recall that $W_Q = (W_Q^1 | \ldots | W_Q^h) = \text{Concat}(W_Q^1, \ldots, W_Q^h)$, and similarly for W_K, W_V. Now notice that $W_Q P_i = W_Q^i$ and we again have the analogous result for W_K, W_V.

Moreover, associativity gives:

Lemma A.1 (Head weights interchange)**.** *For any $X \in \mathbb{R}^{n \times d_{model}}$ and $W \in \mathbb{R}^{d_{model} \times d_{model}}$, the i-th head block of the product equals the product with the i-th head block:*

$$(XW)^i \;=\; XW^i.$$

In particular, this applies to $W = W_Q$, W_K, and W_V, giving $(XW_Q)^i = XW_Q^i$ and analogously for the Key and Value projections.

Proof. By definition of the block decomposition, $(XW)^i = (XW)P_i$. By the observation above, $W^i = WP_i$. Associativity of matrix multiplication gives $(XW)P_i = X(WP_i)$, hence $(XW)^i = XW^i$. □

We can complete the proof of Observation 3.1 here by using the lemma to remark that

$$\begin{aligned}
\text{head}_i(X, W_Q, W_K, W_V) &= \text{softmax}\left(\frac{(XW_Q^i)(XW_K^i)^\top}{\sqrt{d_k}} + M\right)(XW_V^i) \\
&= \text{softmax}\left(\frac{(XW_Q)^i\big((XW_K)^i\big)^\top}{\sqrt{d_k}} + M\right)(XW_V)^i \\
&= g_i(XW_Q, XW_K, XW_V)
\end{aligned}$$

for an appropriate function g_i, and then

$$\begin{aligned}
\mathcal{S}_c(X, W_Q, W_K, W_V) &= \text{Concat}\left(g_1(XW_Q, XW_K, XW_V), \ldots, g_h(XW_Q, XW_K, XW_V)\right) \\
&= g(XW_Q, XW_K, XW_V)
\end{aligned}$$

for the appropriate function g. □

Instead, we make the deliberate choice to introduce a different notation in which this property becomes visible right from the definition of the multi-head attention. The reader can skip the remainder of this section.

Block-wise operations. We decompose matrices into head-wise blocks:

- For $F \in \mathbb{R}^{n \times d_{\text{model}}}$, write $F = (F_1 | \cdots | F_h)$ where each $F_i \in \mathbb{R}^{n \times d_k}$ (e.g. $F = XW_Q$ or $F = XW_V$).
- For $G \in \mathbb{R}^{d_{\text{model}} \times n}$, write $G = \begin{pmatrix} G_1 \\ \vdots \\ G_h \end{pmatrix}$ where each $G_i \in \mathbb{R}^{d_k \times n}$ (e.g. $G = (XW_K)^\top$).

- For $A \in \mathbb{R}^{n\times(hn)}$, write $A = (A_1|\cdots|A_h)$ where each $A_i \in \mathbb{R}^{n\times n}$ (e.g. the matrix of per-head attention logits or weights).

We recall that, in this notation, the standard matrix multiplication of F and G can be written as $FG = \sum_{i=1}^{h} F_i G_i$.

To express per-head operations, we define two *Block Hadamard Products*:

$$F \boxdot^t G = (F_1 G_1|\cdots|F_h G_h) \in \mathbb{R}^{n\times(hn)},$$

$$A \boxdot F = (A_1 F_1|\cdots|A_h F_h) \in \mathbb{R}^{n\times d_{\text{model}}}.$$

These are variants of Hadamard block products introduced in (Horn et al., 1991).

The *block softmax* applies softmax and optional masking independently to each head:

$$\text{Softmax}_M^{\boxdot}(A_1|\cdots|A_h) = \big(\text{softmax}(M + A_1) \,|\, \cdots \,|\, \text{softmax}(M + A_h)\big),$$

where $A_i \in \mathbb{R}^{n\times n}$ are attention logits and $M \in \mathbb{R}^{n\times n}$ is a mask, and we apply the softmax row-wise.

Multi-head attention in block notation. Multi-head self-attention written in the notation given above then reads:

$$\mathcal{S}_c(X, W_Q, W_K, W_V) = \text{Softmax}_M^{\boxdot}\left(\tfrac{1}{\sqrt{d_k}}\,(XW_Q) \boxdot^t (XW_K)^\top\right) \boxdot (XW_V), \qquad (2)$$

$$\text{Attn}(X, W_Q, W_K, W_V, W_O) = \mathcal{S}_c(X, W_Q, W_K, W_V) \cdot W_O.$$

Observation 3.1 now follows directly from the definition (2). All we need to do now is prove that this definition does indeed match the standard one.

Proposition A.2 (Equivalence with the standard definition)**.** *The block-notation expression* (2) *for* $\mathcal{S}_c$ *coincides with the standard concatenation of heads:*

$$\mathcal{S}_c(X, W_Q, W_K, W_V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h).$$

Proof. Denote by $\mathcal{S}_c^{\boxdot}$ the right-hand side of (2). We show that its i-th block equals head_i by tracking the i-th block through each operation in sequence. Recall that

$$\text{head}_i(X, W_Q, W_K, W_V) = \text{softmax}\left(\frac{(XW_Q^i)(XW_K^i)^\top}{\sqrt{d_k}} + M\right)(XW_V^i).$$

Step 1: Logits. By definition of $\boxdot^t$, the product $(XW_Q) \boxdot^t (XW_K)^\top$ concatenates the h blocks $(XW_Q)^i\big((XW_K)^i\big)^\top$ for $i = 1, \ldots, h$. By Lemma A.1, $(XW_Q)^i = XW_Q^i$ and $(XW_K)^i = XW_K^i$, so the i-th block of the scaled logits is

$$\frac{1}{\sqrt{d_k}}\,(XW_Q^i)(XW_K^i)^\top.$$

Step 2: Softmax. By definition, $\text{Softmax}_M^{\boxdot}$ applies the row-wise softmax (with mask M) independently to each $n \times n$ block and concatenates the results. Denoting the output by $\mathcal{A} = (\mathcal{A}_1|\cdots|\mathcal{A}_h)$, the i-th block is

$$\mathcal{A}_i = \text{softmax}\left(\frac{(XW_Q^i)(XW_K^i)^\top}{\sqrt{d_k}} + M\right).$$

Step 3: Value weighting. By definition of $\boxdot$, the final product $\mathcal{A} \boxdot (XW_V)$ multiplies the i-th $n \times n$ block of $\mathcal{A}$ with the i-th $n \times d_k$ block of (XW_V) and concatenates the results:

$$\mathcal{A} \boxdot (XW_V) = \big(\mathcal{A}_1(XW_V)^1 \,\big|\, \cdots \,\big|\, \mathcal{A}_h(XW_V)^h\big).$$

By Lemma A.1, $(XW_V)^i = XW_V^i$, so the i-th block of $\mathcal{S}_c^{\boxdot}$ is

$$\left[\mathcal{S}_c^{\boxdot}\right]_i = \mathcal{A}_i \cdot XW_V^i = \mathrm{softmax}\left(\frac{(XW_Q^i)(XW_K^i)^\top}{\sqrt{d_k}} + M\right)(XW_V^i) = \mathrm{head}_i(X, W_Q, W_K, W_V).$$

Conclusion. Since $\mathcal{S}_c^{\boxdot}$ and $\mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)$ are both concatenations of the same h blocks, they are equal. □

Proof of Observation 3.1. By Proposition A.2, the block-notation formula (2) is an equivalent expression for $\mathcal{S}_c$. Setting $A = XW_Q$, $B = XW_K$, $C = XW_V$, it reads

$$\mathcal{S}_c(X, W_Q, W_K, W_V) = \mathrm{Softmax}_M^{\boxdot}\left(\tfrac{1}{\sqrt{d_k}}\, A \,\boxdot^t\, B^\top\right) \boxdot C \;=:\; g(A, B, C).$$

The right-hand side depends on X only through the products A, B, C. □

We finish this section by providing the proof of Proposition 3.2: block-diagonal transformations cancel head-by-head inside $\boxdot^t$, leaving $\mathcal{S}_c$ unchanged.

Proof of Proposition 3.2. Since $D = \mathrm{diag}(D_1, \ldots, D_h)$ is block-diagonal, the i-th head block of $W_Q D$ is $W_Q^i D_i$ and the i-th head block of $W_K(D^\top)^{-1}$ is $W_K^i(D_i^\top)^{-1}$. Therefore, the i-th block of the attention logits satisfies:

$$(XW_Q^i D_i)\big(XW_K^i(D_i^\top)^{-1}\big)^\top = XW_Q^i\, D_i\, D_i^{-1}\, (XW_K^i)^\top = XW_Q^i\, (XW_K^i)^\top,$$

where we used $\big((D_i^\top)^{-1}\big)^\top = D_i^{-1}$. Since this holds for every head, $(XW_QD) \boxdot^t (XW_K(D^\top)^{-1})^\top = (XW_Q) \boxdot^t (XW_K)^\top$. The value projection W_V is unchanged, so $\mathcal{S}_c(X,\, W_QD,\, W_K(D^\top)^{-1},\, W_V) = \mathcal{S}_c(X,\, W_Q,\, W_K,\, W_V)$. □

B Attention Scaling Derivation

We derive the corrective scaling factor $\frac{1}{2\sqrt{d_k}}$ for the reduced architecture, even though we believe that this constant should ultimately be handled empirically. Let $x \in \mathbb{R}^{d_{\text{model}}}$ be the input to the attention layer. For the simplicity of the derivation, we may assume that the input rows have unit norm, $\|x\|^2 = 1$. Weights are initialized from $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.02$.

Standard Baseline ($q_i = xW_Q^i$). The query for head i is a linear projection where $W_Q^i \in \mathbb{R}^{d_{\text{model}} \times d_k}$. Since each entry of $W_Q^i \sim \mathcal{N}(0, \sigma^2)$, the variance of each component of q_i is $\|x\|^2\sigma^2 = \sigma^2$. The attention score $S_{\text{std}} = q_i \cdot (xW_K^i)^\top$ is the dot product of two d_k-dimensional vectors with component variance σ^2, yielding an initial standard deviation of:

$$\mathrm{StdDev}(S_{\text{std}}) = \sigma^2\sqrt{d_k} = (0.02)^2\sqrt{64} = 0.0032.$$

Reduced Architecture ($W_Q = I$). The query for head i is defined as a coordinate slice of the input: $q_i = x_{[i\cdot d_k:(i+1)\cdot d_k]}$. Because the total norm $\|x\|^2 = 1$ is distributed across the h heads, the expected squared norm of this slice is $\|q_i\|^2 = \frac{1}{h}$. The attention score $S_{\text{red}} = q_i \cdot (xW_K^i)^\top$ has a variance that depends only on the key weight variance:

$$\mathrm{Var}(S_{\text{red}}) = \|q_i\|^2\sigma^2 = \frac{\sigma^2}{h} \implies \mathrm{StdDev}(S_{\text{red}}) = \frac{\sigma}{\sqrt{h}} = \frac{0.02}{\sqrt{12}} \approx 0.0058.$$

Corrective Factor. Comparing the two regimes, the initial scores in the reduced architecture are approximately $1.8\times$ larger than the baseline ($\frac{0.0058}{0.0032} \approx 1.8$). To prevent early softmax saturation and maintain a starting variance consistent with standard transformers, we introduce a corrective factor to the scaling. Since $1/1.8 \approx 0.55$, we adopt the factor $\frac{1}{2}$ as a clean and effective approximation, resulting in the modified scaling $\frac{1}{2\sqrt{d_k}}$.

C LayerNorm and Basis Transformations

In Appendices C–E, we adopt the standard mathematical convention: vectors are column vectors and linear maps act by left multiplication (Ax), in contrast to the ML convention (xA) used in the main text.

We investigate when basis transformations commute with LayerNorm. Define $L_\varepsilon : \mathbb{R}^d \to \mathbb{R}^d$ by

$$L_\varepsilon(x) = \frac{x - \mu(x)\mathbf{1}}{\sigma_\varepsilon(x)},$$

where $\mu(x) = \frac{1}{d}\sum_{i=1}^d x_i$ and $\sigma_\varepsilon(x) = \sqrt{\frac{1}{d}\sum_{i=1}^d (x_i - \mu(x))^2 + \varepsilon}$.

Lemma C.1 (LayerNorm Semi-Conjugacy Lemma)**.** *Let $\varepsilon > 0$ and $A \in \mathrm{GL}(d)$ be fixed. Then there exists a function $f : \mathbb{R}^d \to \mathbb{R}^d$ and a diagonal matrix $D \in \mathbb{R}^{d\times d}$ such that*

$$L_\varepsilon(f(x)) = (DA)L_\varepsilon(x) \quad \forall x \in \mathbb{R}^d.$$

Here, D corresponds to the learned per-coordinate scaling parameter of LayerNorm; our construction explicitly leverages this degree of freedom to enable the basis transformation. Without the scaling parameter, this result would not hold in general. Moreover, for any scalar field $h : \mathbb{R}^d \to \mathbb{R}$, $f + h\mathbf{1}$ also satisfies this equation.

Proof. Let $H = \{z \in \mathbb{R}^d \mid \mathbf{1}^\top z = 0\}$ be the zero-mean subspace. Since $\mu(L_\varepsilon(x)) = 0$, the image of L_ε lies in H.

Step 1: Constructing the equivariant matrix M_0. We require $M = DA$ to preserve H. Let $\tilde{v} = (A^\top)^{-1}\mathbf{1}$ and $v = \tilde{v}/\|\tilde{v}\|$. Define $D_\lambda = \lambda \cdot \mathrm{diag}(v)$. For any $z \in H$:

$$\mathbf{1}^\top(D_\lambda A)z = \lambda v^\top Az = \frac{\lambda}{\|\tilde{v}\|}(A^\top\tilde{v})^\top z = \frac{\lambda}{\|\tilde{v}\|}\mathbf{1}^\top z = 0,$$

where we used $A^\top\tilde{v} = \mathbf{1}$ by definition of $\tilde{v}$. Thus $(D_\lambda A)(H) \subset H$. Choose λ_0 such that the restriction $\|M_{0|H}\|_2 = 1$ where $M_0 = D_{\lambda_0}A$.

Step 2: Domain and unique inverse. For any x, we have $\|L_\varepsilon(x)\|^2 = d\sigma_0(x)^2/(\sigma_0(x)^2 + \varepsilon) < d$, so the image of L_ε is the open ball $B = \{z \in H \mid \|z\| < \sqrt{d}\}$. Since $\|M_{0|H}\|_2 = 1$, we have $M_0(B) \subset B$. For $z \in B$, solving $z = L_\varepsilon(c)$ for $c \in H$ (i.e., $z = c/\sqrt{\|c\|^2/d + \varepsilon}$) gives $\|c\| = \|z\|\sqrt{\varepsilon}/\sqrt{d - \|z\|^2}$, hence:

$$L_\varepsilon^{-1}(z) = \sqrt{\frac{d\varepsilon}{d - \|z\|^2}}\, z.$$

Step 3: Defining f. Set $f(x) = L_\varepsilon^{-1}(M_0 L_\varepsilon(x)) + h(x)\mathbf{1}$ for any scalar field h. Then:

$$L_\varepsilon(f(x)) = L_\varepsilon(L_\varepsilon^{-1}(M_0 L_\varepsilon(x))) = M_0 L_\varepsilon(x) = (DA)L_\varepsilon(x). \qquad \square$$

Remark 5 (An informal discussion on the approximate linearity of f_M in high dimensions)**.** *Lemma C.1 relies on a carefully constructed matrix $M_0 = D_{\lambda_0}A$ to establish an algebraic semi-conjugacy. We examine how $f_M(x) = L_\varepsilon^{-1}(ML_\varepsilon(x))$ behaves for a generic matrix M in high dimensions.*

SVD reduction. *On the zero-mean subspace H, the normalization L_ε is a radial map of the form $x \mapsto \phi(\|x\|) \cdot x$. Since the scaling depends only on $\|x\|$, it commutes with any orthogonal transformation that preserves H. If $M = U\Sigma V^T$ is the SVD, the orthogonal factors pass through the nonlinearity exactly: $f_M(x) = U \cdot f_\Sigma(V^T x)$. The rotations can be absorbed into surrounding weight matrices. The entire nonlinear content of f_M reduces to the diagonal map $f_\Sigma(y) = \alpha(y) \cdot \Sigma y$, where $\alpha(y) = (1 + \Delta(y)/(d\varepsilon))^{-1/2}$ and $\Delta(y) = \|y\|^2 - \|\Sigma y\|^2$ is the metric defect. The scalar α equals 1 exactly when Σ is an isometry.*

Linearization. *At the origin, $Df_\Sigma(0) = \Sigma$, so the linearization is exact and the quality of the approximation is controlled by $|1 - \alpha(y)|$. For small inputs ($\|y\|^2 \ll d\varepsilon$), the ε-smoothing forces*

$\alpha \approx 1$ *regardless of* M*'s spectrum, extending the linearization to a ball of radius* $\mathcal{O}(\sqrt{d\varepsilon})$*. For large inputs, the linearization breaks down unless* Σ *is close to an isometry.*

The stochastic picture. *For a random matrix* M *with i.i.d.* $\mathcal{N}(0, 1/d)$ *entries, the metric defect self-averages by concentration of measure (Vershynin, 2018). The mean defect is* $\mathbb{E}[\Delta(y)] = \|y\|^2(1 - Tr(\Sigma^2)/d)$*, which is small for square matrices at variance* $1/d$*. By standard concentration on the sphere,* $|1 - \alpha(y)| = \mathcal{O}(\|y\|^2/(d^{3/2}\varepsilon))$*, extending the linearization well beyond the deterministic small-input regime. Provided* $\varepsilon \gg \|y\|^2 d^{-3/2}$*,* $f_M(x) \approx Mx$ *with quality improving as* d *grows.*

Remark 6 (Extension to RMSNorm and Dynamic Tanh)**.** *The argument above extends to RMSNorm (Zhang & Sennrich, 2019), which has become the normalization of choice in many modern LLMs (Touvron et al., 2023). Furthermore, a similar reasoning applies to the recently proposed Dynamic Tanh (DyT) alternative to normalization (Zhu et al., 2025). Indeed, consider the map*

$$DyT_{\gamma,\alpha,\beta} : \mathbb{R}^d \to \mathbb{R}^d, \quad (x_1, \ldots, x_d) \mapsto diag(\gamma)\big(\tanh(\alpha_1 x_1), \ldots, \tanh(\alpha_d x_d)\big) + \beta.$$

This transformation is bijective onto its image whenever $\gamma_i \neq 0$ *and* $\alpha_i \neq 0$ *for all* i*, a condition that is reasonable due to the stochasticity of optimization. Hence, the structural result established above extends beyond LayerNorm to both RMSNorm and Dynamic Tanh.*

Theorem C.2 (Basis Transformation Through LayerNorm)**.** *Let* $\Theta \in \mathrm{GL}(d)$ *be a basis transformation,* $\mathrm{MLP} : \mathbb{R}^d \to \mathbb{R}^d$ *a multilayer perceptron, and* D *the learned LayerNorm scaling. If there exist* MLP' *and* D' *such that*

$$D' L_\varepsilon(x + \mathrm{MLP}'(x)) = \Theta D L_\varepsilon(x + \mathrm{MLP}(x)) \quad \forall x,$$

then by Lemma C.1, MLP' *must satisfy:*

$$\mathrm{MLP}'(x) = L_\varepsilon^{-1}(M_0 L_\varepsilon(x + \mathrm{MLP}(x))) - x + h(x)\mathbf{1}$$

where M_0 *is constructed as above for* $A = \Theta D$*.*

Remark 7 (Theoretical obstruction to exact Query weight elimination with LayerNorm)**.** *This theorem reveals that preserving a basis transformation through LayerNorm requires* MLP' *to approximate a highly nonlinear function involving the composition of normalization, the original MLP, and denormalization. Since standard MLPs of the form* $W_2\sigma(W_1 x)$ *cannot exactly represent this composition, exact Query weight elimination with LayerNorm is not achievable using standard architectural components. This theoretical obstruction motivates either removing LayerNorm or accepting approximate equivalence with adjusted hyperparameters, as we pursue in our experiments. Notably, we test both a slightly larger hidden dimension in the MLP, as well as reduced weight decay compared to the original model.*

D MLP Basis Transfer Approximation

We continue with the column-vector convention (see Appendix C).

Standard transformers include skip connections around both attention and MLP blocks. With MLP skip connections, the basis transformation leads to the functional equation:

$$\Omega_1 \circ (\mathrm{MLP} + \mathrm{Id}) \circ \Omega_2^{-1} \approx \mathrm{MLP}' + \mathrm{Id} \tag{3}$$

We test whether gradient descent can discover approximate solutions.

Setup. We generate synthetic targets $Y_{\text{target}}(X) = W_2 \cdot \mathrm{GELU}(W_1 X) + ZX - X$ with random Gaussian weights, and train a GELU MLP with skip connection $Y_{\text{model}}(X) = W_2' \cdot \mathrm{GELU}(W_1' X) + X$ to approximate it. We use AdamW for 20,000 steps with batch size 65,536, testing dimensions $h \in \{256, 512, 768\}$.

Results. Figure 2 shows the trained MLP achieves 4 to 6% relative L2 error across all dimensions, substantially outperforming the optimal linear baseline (11 to 14%). Mean cosine similarity reaches 0.999 ($\approx 2.5^\circ$ angular error). This demonstrates that MLPs can implicitly learn basis adaptations through gradient descent, bridging the gap between our theoretical guarantees (exact solutions, restricted skip connections) and practical architectures (approximate solutions, full skip connections).

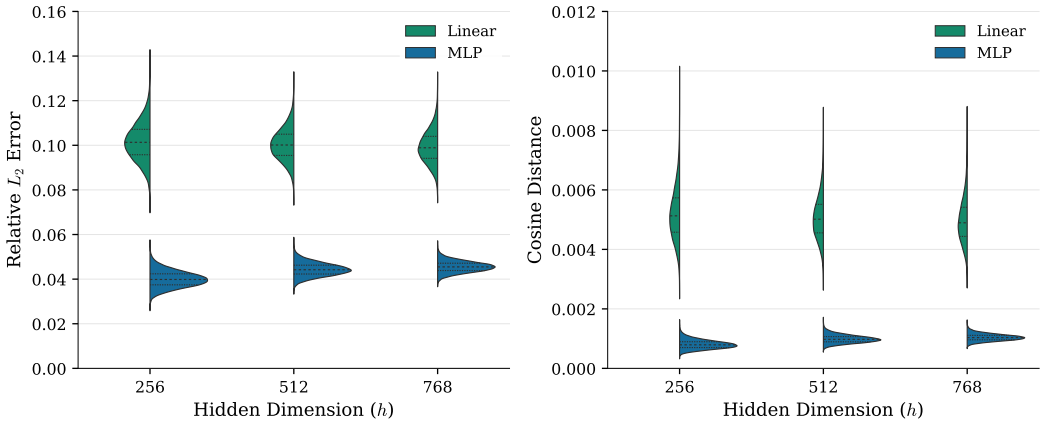

Figure 2: **Left:** Relative L_2 error for approximating basis-transformed skip-connected MLPs as a function of hidden dimension $h \in \{256, 512, 768\}$. The trained GELU MLP achieves relative error between 4% and 6%, significantly lower than the 11% to 14% error of the optimal linear baseline. **Right:** Mean cosine distance between predicted and target outputs. The MLP maintains a distance of approximately 0.002 across all dimensions, while the linear baseline remains above 0.009, indicating superior directional alignment.

E When Can Skip Connections Be Absorbed into ReLU MLPs?

We continue with the column-vector convention (see Appendix C).

A fundamental question in deep learning theory is understanding when architectural components like skip connections are *expressively necessary* versus merely helpful for optimization. ReLU networks compute continuous piecewise linear (CPWL) functions, partitioning input space into convex polyhedra, each associated with a distinct affine mapping, via "bent" hyperplane arrangements (Grigsby & Lindsey, 2022). The geometry of these arrangements determines the network's decision boundaries and expressivity (Arora et al., 2018). The number of linear regions serves as a coarse measure of expressivity, growing polynomially in width but exponentially in depth (Montufar et al., 2014; Telgarsky, 2016). While depth separation results establish that deep networks can represent functions requiring exponentially many neurons in shallow networks, the role of skip connections in *expressivity* remains less understood. We provide an exact algebraic characterization of when skip connections can be absorbed into ReLU MLPs *without increasing width*, revealing that skip and non-skip architectures access fundamentally different regions of function space.

Theorem E.1 (Residual-free representability of skip-ReLU-MLPs)**.** *Let $h \geq 2$ and set $m = 4h$. Let*

$$W_1 \in \mathbb{R}^{m \times h}, \qquad W_2 \in \mathbb{R}^{h \times m}$$

be given matrices with $\mathrm{rank} W_1 = \mathrm{rank} W_2 = h$, where W_1 consists of pairwise non-collinear, non-zero rows, and W_2 consists of non-zero columns. Denote ReLU coordinatewise. There exist matrices

$$V_1 \in \mathbb{R}^{m \times h}, \qquad V_2 \in \mathbb{R}^{h \times m}$$

such that the identity

$$W_2 \,\mathrm{ReLU}(W_1 x) + x \;=\; V_2 \,\mathrm{ReLU}(V_1 x) \qquad \textit{holds for all } x \in \mathbb{R}^h$$

if and only if there exists an index set $J \subset \{1, \ldots, m\}$ with $|J| \geq h$ for which

$$W_2[:, J]\, W_1[J, :] = -I_h.$$

Proof. We prove both directions.

(Sufficiency). Suppose $J \subset \{1, \ldots, m\}$ satisfies $|J| \geq h$ and $W_2[:, J]\, W_1[J, :] = -I_h$. Let $\Pi = \mathrm{diag}(\mathbf{1}_J) \in \{0, 1\}^{m \times m}$ be the coordinate projector onto J, and set the sign diagonal $D := I - 2\Pi \in$

$\{\pm 1\}^{m\times m}$. Define $V_1 := DW_1$ and $V_2 := W_2$. For any $y \in \mathbb{R}^m$ write $\mathrm{ReLU}(y) = \frac{1}{2}(y + |y|)$. Since D has diagonal entries ± 1 we have $|V_1 x| = |W_1 x|$ for all x. Hence

$$V_2\mathrm{ReLU}(V_1x) = \tfrac{1}{2}V_2\big(V_1x + |V_1x|\big) = \tfrac{1}{2}\big(V_2V_1x + V_2|W_1x|\big),$$
$$W_2\mathrm{ReLU}(W_1x) + x = \tfrac{1}{2}\big(W_2W_1x + W_2|W_1x| + 2x\big).$$

Using $D = I - 2\Pi$ and $W_2\Pi W_1 = -I_h$ we obtain $V_2V_1 = W_2DW_1 = W_2(I-2\Pi)W_1 = W_2W_1 + 2I_h$. Substituting into the displayed expressions yields $V_2\mathrm{ReLU}(V_1x) = W_2\mathrm{ReLU}(W_1x) + x$ for all x.

(Necessity). Suppose the identity $W_2\,\mathrm{ReLU}(W_1x) + x = V_2\,\mathrm{ReLU}(V_1x)$ holds for all $x \in \mathbb{R}^h$. Using $\mathrm{ReLU}(y) = \frac{1}{2}(y + |y|)$, and separating the odd and even terms in x yields: $(W_2W_1 + 2I_h - V_2V_1)x = V_2|V_1x| - W_2|W_1x|$. The left-hand side is odd, while the right-hand side is even. This implies both sides must be identically zero, giving: (N1) $V_2V_1 = W_2W_1 + 2I_h$ and (N2) $V_2|V_1x| = W_2|W_1x|$ for all x.

We now analyze condition (N2) to relate the rows of V_1 and W_1. The function $f(x) = W_2|W_1x|$ is non-differentiable exactly on the set $\mathcal{H}_W = \bigcup_{i:(W_2)_{:i}\neq 0}\{x \mid (W_1)_i x = 0\}$. Since W_2 consists of non-zero columns, every hyperplane corresponding to a row of W_1 is present in the singular set. Since W_1 consists of pairwise non-collinear rows, these hyperplanes are distinct. For the identity (N2) to hold, the set of non-differentiable points must be identical: $\mathcal{H}_W = \mathcal{H}_V$. This implies that for every row w of W_1, there exists a row v of V_1 such that they define the same hyperplane (i.e., they are collinear). Thus, V_1 must be a scaled permutation of W_1. We can write $V_1 = DPW_1$ for some diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_m)$ with $d_i \neq 0$ and some permutation matrix P.

Substitute $V_1 = DPW_1$ into (N2). Let $D_a = \mathrm{diag}(|d_1|, \ldots, |d_m|)$. Since $|Py| = P|y|$, we get $V_2D_aP|W_1x| = W_2|W_1x|$. Since $h \geq 2$ and the rows of W_1 are pairwise non-collinear, this forces $V_2D_aP = W_2$, giving $V_2 = W_2P^TD_a^{-1}$.

Substituting into (N1): $W_2P^T(D_a^{-1}D)PW_1 = W_2W_1 + 2I_h$. Let $S = D_a^{-1}D$ (entries ± 1) and $D' = P^TSP$. Then $W_2(D' - I)W_1 = 2I_h$, giving $W_2(-2\Pi)W_1 = 2I_h$ where Π projects onto indices where D' has -1 entries. Thus $W_2[:, J]W_1[J, :] = -I_h$.

By Sylvester's inequality, $h = \mathrm{rank}(-I_h) \leq |J|$, so $|J| \geq h$. □

Corollary E.2 (Uniqueness under invertible residual perturbation)**.** *Let $h \geq 2$ and $m = 4h$. For matrices $W_1 \in \mathbb{R}^{m\times h}$, $W_2 \in \mathbb{R}^{h\times m}$, and $Z \in \mathbb{R}^{h\times h}$, define $\phi_{W_1,W_2,Z}(x) = W_2\,\mathrm{ReLU}(W_1x) + Zx$.*

Consider the parameter space $\mathcal{W} := \mathbb{R}^{m\times h} \times \mathbb{R}^{h\times m} \times \mathbb{R}^{h\times h}$. For Lebesgue-almost every $(W_1, W_2, Z) \in \mathcal{W}$, the matrices satisfy the conditions of Theorem E.1 (pairwise non-collinear non-zero rows for W_1, non-zero columns for W_2), and the following holds:

If $Z' \neq Z$ with $Z - Z'$ invertible, then for every (W_1', W_2'): $\phi_{W_1,W_2,Z} \neq \phi_{W_1',W_2',Z'}$.

Remark 8. *Theorem E.1 and Corollary E.2 hold without modification for any $m \geq h$.*

Proof. Suppose for contradiction that $\phi_{W_1,W_2,Z} = \phi_{W_1',W_2',Z'}$. Then for all $x \in \mathbb{R}^h$:

$$W_2\mathrm{ReLU}(W_1x) + Zx = W_2'\mathrm{ReLU}(W_1'x) + Z'x.$$

Rearranging:

$$W_2\mathrm{ReLU}(W_1x) + (Z - Z')x = W_2'\mathrm{ReLU}(W_1'x).$$

Since $Z - Z'$ is invertible, substituting $y = (Z - Z')x$ yields:

$$W_2\mathrm{ReLU}(W_1(Z - Z')^{-1}y) + y = W_2'\mathrm{ReLU}(W_1'(Z - Z')^{-1}y).$$

Setting $A = W_2$ and $B = W_1(Z - Z')^{-1}$, the left-hand side has the form $A\mathrm{ReLU}(By) + y$. By the necessity direction of Theorem E.1, there must exist J with $|J| \geq h$ such that $A[:, J]B[J, :] = -I_h$. Substituting back and right-multiplying by $Z - Z'$:

$$W_2[:, J]W_1[J, :] = Z' - Z.$$

For any fixed J, this imposes h^2 algebraic constraints on (W_1, W_2, Z). Under continuous distributions, this set has Lebesgue measure zero. Since there are finitely many choices of J, the union over all J also has measure zero, so for almost every (W_1, W_2, Z) no such J exists. □

The condition $W_2[:, J]W_1[J, :] = -I_h$ is *non-generic*: it fails with probability 1 under continuous weight distributions. This reveals that ReLU MLPs and Skip-ReLU-MLPs of the same width represent *generically disjoint* function classes. The algebraic condition characterizes exactly when these classes intersect. This clarifies the role of skip connections: they do not provide strictly "more" expressivity, but rather access to a *different* region of function space.

F Proofs from Sections 3 and 4

Proof of Lemma 3.1 (Reparametrization Lemma). Let $\Theta = W_Q$, $\widetilde{W}_K = W_Q^{-1}W_K$, and $\widetilde{W}_V = W_Q^{-1}W_V$. Then for any $X \in \mathrm{Mat}(n, d)$:

$$\begin{aligned} f(X\Theta, I_d, \widetilde{W}_K, \widetilde{W}_V) &= g(X\Theta \cdot I_d, X\Theta \cdot \widetilde{W}_K, X\Theta \cdot \widetilde{W}_V) \\ &= g(XW_Q, XW_Q \cdot W_Q^{-1}W_K, XW_Q \cdot W_Q^{-1}W_V) \\ &= g(XW_Q, XW_K, XW_V) = f(X, W_Q, W_K, W_V). \end{aligned}$$ □

Proof of Theorem 4.2 (Single-Layer Free Lunch). Let $\Theta = W_Q^j$ for the target layer j. The block structure with all skip connections is:

$$Y_i = X_{i-1} + \mathrm{Attn}_i(X_{i-1}), \qquad X_i = Y_i + \mathrm{MLP}_i(Y_i).$$

Inductive hypothesis $\mathcal{H}(i)$: The output of layer i in the modified model is $\tilde{X}_i = X_i\Theta$, where X_i is the original output.

Base case $\mathcal{H}(0)$: Set $\tilde{E} = E\Theta$ and $\tilde{E}_P = E_P\Theta$. Then $\tilde{X}_0 = X_0\Theta$, so $\mathcal{H}(0)$ holds.

Inductive step $\mathcal{H}(i-1) \Rightarrow \mathcal{H}(i)$: Assume $\tilde{X}_{i-1} = X_{i-1}\Theta$. Define:

$$\begin{aligned} &\tilde{W}_Q^i = \Theta^{-1}W_Q^i, && \tilde{W}_K^i = \Theta^{-1}W_K^i, && \tilde{W}_V^i = \Theta^{-1}W_V^i, \\ &\tilde{W}_O^i = W_O^i\Theta, && \tilde{W}_u^i = \Theta^{-1}W_u^i, && \tilde{W}_d^i = W_d^i\Theta. \end{aligned}$$

For attention, the projections are:

$$\begin{aligned} \text{Query:} \quad & (X_{i-1}\Theta)(\Theta^{-1}W_Q^i) = X_{i-1}W_Q^i, \\ \text{Key:} \quad & (X_{i-1}\Theta)(\Theta^{-1}W_K^i) = X_{i-1}W_K^i, \\ \text{Value:} \quad & (X_{i-1}\Theta)(\Theta^{-1}W_V^i) = X_{i-1}W_V^i. \end{aligned}$$

Thus $\mathcal{S}_c(X_{i-1}\Theta, \tilde{W}_Q^i, \tilde{W}_K^i, \tilde{W}_V^i) = \mathcal{S}_c(X_{i-1}, W_Q^i, W_K^i, W_V^i)$, and the attention output is:

$$\begin{aligned} \widetilde{\mathrm{Attn}}_i(X_{i-1}\Theta) &= \mathcal{S}_c(X_{i-1}\Theta, \tilde{W}_Q^i, \tilde{W}_K^i, \tilde{W}_V^i) \cdot \tilde{W}_O^i \\ &= \mathcal{S}_c(X_{i-1}, W_Q^i, W_K^i, W_V^i) \cdot W_O^i\Theta \\ &= \mathrm{Attn}_i(X_{i-1}) \cdot \Theta. \end{aligned}$$

After the attention skip connection:

$$\tilde{Y}_i = X_{i-1}\Theta + \widetilde{\mathrm{Attn}}_i(X_{i-1}\Theta) = (X_{i-1} + \mathrm{Attn}_i(X_{i-1}))\Theta = Y_i\Theta.$$

For the MLP:

$$\widetilde{\mathrm{MLP}}_i(Y_i\Theta) = \phi(Y_i\Theta \cdot \Theta^{-1}W_u^i)W_d^i\Theta = \phi(Y_iW_u^i)W_d^i\Theta = \mathrm{MLP}_i(Y_i)\Theta.$$

After the MLP skip connection:

$$\tilde{X}_i = \tilde{Y}_i + \widetilde{\mathrm{MLP}}_i(\tilde{Y}_i) = Y_i\Theta + \mathrm{MLP}_i(Y_i)\Theta = X_i\Theta.$$

Thus $\mathcal{H}(i)$ holds.

Key observation for layer j: Since $\Theta = W_Q^j$:

$$\tilde{W}_Q^j = \Theta^{-1}W_Q^j = (W_Q^j)^{-1}W_Q^j = I_d.$$

LM head: Set $\tilde{W}_{\mathrm{LM}} = \Theta^{-1}W_{\mathrm{LM}}$. Then $X_L\Theta \cdot \Theta^{-1}W_{\mathrm{LM}} = X_LW_{\mathrm{LM}}$. □

Proof of Theorem 4.3 (Weight-Shared Query Weight Elimination). We first establish that blocks are conjugate to reduced blocks.

> **Lemma (Blocks are linear conjugates of Reduced blocks).** Let $B : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ be a transformer block without normalization given by $B(X) = (\mathrm{Id} + \mathrm{MLP}) \circ (\mathrm{Id} + \mathrm{Attn})(X)$ where Attn uses weights (W_Q, W_K, W_V, W_O) with $W_Q \in \mathrm{GL}(d)$, and $\mathrm{MLP}(Y) = \phi(YW_{\text{up}})W_{\text{down}}$ for an element-wise activation ϕ. Then there exists $\Theta \in \mathrm{GL}(d)$ and a reduced block $\widetilde{B}$ using weights $(\mathrm{Id}, \widetilde{W}_K, \widetilde{W}_V, \widetilde{W}_O, \widetilde{W}_{\text{up}}, \widetilde{W}_{\text{down}})$ such that
>
> $$B = \Theta^{-1} \circ \widetilde{B} \circ \Theta.$$
>
> *Proof of Lemma.* Set $\Theta = W_Q$, $\widetilde{W}_K = \Theta^{-1}W_K$, $\widetilde{W}_V = \Theta^{-1}W_V$, $\widetilde{W}_O = W_O\Theta$, $\widetilde{W}_{\text{up}} = \Theta^{-1}W_{\text{up}}$, and $\widetilde{W}_{\text{down}} = W_{\text{down}}\Theta$. Direct computation verifies the conjugacy relation. □

Since all layers share parameters, applying the lemma yields $B^{\circ L} = (\Theta^{-1} \circ \widetilde{B} \circ \Theta)^{\circ L} = \Theta^{-1} \circ \widetilde{B}^{\circ L} \circ \Theta$. Setting $\widetilde{E} = E\Theta$, $\widetilde{E}_P = E_P\Theta$, and $\widetilde{W}_{\text{LM}} = \Theta^{-1}W_{\text{LM}}$ completes the equivalence. □

Proof of Theorem 4.4 (Attention-Skip-Only Query Weight Elimination). We first state the key structural result.

> **Lemma (Attention-Skip-Only Block Intertwining).** Let $B_i : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ be the i-th block with weights $(W_Q^i, W_K^i, W_V^i, W_O^i, W_{\mathrm{MLP}_u}^i, W_{\mathrm{MLP}_d}^i)$, skip around attention only:
>
> $$B_i(X) = \phi\big((X + \mathcal{S}_c(X, W_Q^i, W_K^i, W_V^i)W_O^i) \cdot W_{\mathrm{MLP}_u}^i\big)W_{\mathrm{MLP}_d}^i$$
>
> Then for $\Theta_i = W_Q^i$ and any $\Theta_{i+1} \in \mathrm{GL}(d)$, there exists a reduced block $\widetilde{B}_i$ with $\widetilde{W}_Q^i = I_d$ such that:
>
> $$B_i(X) = \widetilde{B}_i(X \cdot \Theta_i) \cdot \Theta_{i+1}^{-1}$$
>
> The reduced weights are: $\widetilde{W}_K^i = \Theta_i^{-1}W_K^i$, $\widetilde{W}_V^i = \Theta_i^{-1}W_V^i$, $\widetilde{W}_O^i = W_O^i\Theta_i$, $\widetilde{W}_{\mathrm{MLP}_u}^i = \Theta_i^{-1}W_{\mathrm{MLP}_u}^i$, $\widetilde{W}_{\mathrm{MLP}_d}^i = W_{\mathrm{MLP}_d}^i\Theta_{i+1}$.
>
> *Proof of Lemma.* The attention output with input $X\Theta_i$ is $\widetilde{\mathrm{Attn}}_i(X\Theta_i) = \mathcal{S}_c(X\Theta_i, I_d, \Theta_i^{-1}W_K^i, \Theta_i^{-1}W_V^i)W_O^i\Theta_i = \mathrm{Attn}_i(X)\Theta_i$ by the Reparametrization Lemma. The attention-plus-skip becomes $(X + \mathrm{Attn}_i(X))\Theta_i = Z_i\Theta_i$. The MLP output is $\phi(Z_i\Theta_i \cdot \Theta_i^{-1}W_{\mathrm{MLP}_u}^i)W_{\mathrm{MLP}_d}^i\Theta_{i+1} = B_i(X)\Theta_{i+1}$. □

Define $\Theta_i = W_Q^i$ for $1 \le i \le L$, and $\Theta_{L+1} = I_d$ (untied) or $\Theta_{L+1} = (\Theta_1^\top)^{-1}$ (tied).

Inductive hypothesis $\mathcal{H}(i)$: The reduced model's output after block i equals $X_i \cdot \Theta_{i+1}$.

Base case $\mathcal{H}(0)$: Set $\widetilde{E} = E\Theta_1$, $\widetilde{E}_P = E_P\Theta_1$. The input to block 1 is $\widetilde{X}_0 = X_0\Theta_1$, so $\mathcal{H}(0)$ holds.

Inductive step $\mathcal{H}(i-1) \Rightarrow \mathcal{H}(i)$: By $\mathcal{H}(i-1)$, the input to block i is $X_{i-1}\Theta_i$. By the Intertwining Lemma: $\widetilde{B}_i(X_{i-1}\Theta_i) = B_i(X_{i-1})\Theta_{i+1} = X_i\Theta_{i+1}$.

LM Head: For untied weights, $\Theta_{L+1} = I_d$ gives output X_LW_{LM}. For tied weights, $\widetilde{W}_{\text{LM}} = \Theta_1^\top E^\top$ gives $X_L\Theta_{L+1}\widetilde{W}_{\text{LM}} = X_LW_{\text{LM}}$. □