W_K, W_V is (Linearly) All You Need: On the Necessity of the QKV Weight Triplet in Self-Attention Transformers

Marko Karbevski [1] **Antonij Mijoski** [2]

Abstract

Multi-head attention is invariant under the joint $\mathrm{GL}(d)$ action $(X, W_Q, W_K, W_V) \mapsto (X\Theta,\, \Theta^{-1}W_Q,\, \Theta^{-1}W_K,\, \Theta^{-1}W_V)$, with two consequences for the QKV triplet. First, any one of W_Q, W_K, W_V can be fixed to I_d without loss of expressivity if the corresponding product XW is precomputed; under mild structural conditions the precomputation folds into the preceding MLP (or, in the first layer, the embedding) at no parameter cost, removing 25% of attention parameters per layer. We prove the multi-layer reduction and analyse the practical obstructions. Second, every linear W_Q already lies on the orbit of I_d, so a learned linear query is redundant: expressive gains in the QKV pathway require at least one of the three to be nonlinear, a branch we realise with the residual query $Q(X) = \frac{1}{2}(X + f_\theta(X))$ at parity of parameters. This research also led us to examine residual skip connections: MLPs with and without a skip form generically disjoint function classes for modern activations. We validate both halves on GPT-style models trained from scratch under batch-matched comparisons: the reduced 117M model matches the dense 124M baseline; reallocating the saved parameters to the feed-forward sublayer strictly improves on it; and the residual nonlinear query yields a 2.40% relative log-loss improvement, while a model with 12.5% more non-embedding parameters achieves 0.94%: a $2.5\times$ relative gain.

[1]In Simplicity Technologies [2]IRMA, Université de Strasbourg, France. Correspondence to: Marko Karbevski <marko.karbevski@insimplicity.tech, marko.karbevski@gmail.com>, Antonij Mijoski <antonij.mijoski@unistra.fr, antonijmijo@gmail.com>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

This paper is based on the following articles, where proofs in full generality and the most up-to-date versions of this work can be found: "W_K, W_V is Probably All You Need" (arXiv:2510.23912), "Beyond Linearity in Attention Projections" (arXiv:2603.13381), and "Can an MLP Absorb Its Own Skip Connection?" (arXiv:2604.23705).

1 Introduction

The transformer's multi-head attention (Vaswani et al., 2017) depends on the input X only through the three products XW_Q, XW_K, XW_V. This shared dependence induces a continuous symmetry: any invertible linear change of basis in the ambient space can be absorbed into the three projections, leaving the layer output invariant. The symmetry group is $\mathrm{GL}(d)$, acting jointly on the input-weight tuple (X, W_Q, W_K, W_V).

Two consequences follow directly. **(i) Parameter redundancy.** Because W_Q is almost surely invertible, the choice $\Theta = W_Q$ fixes $W_Q \mapsto I_d$ without altering the layer's output, provided the input to the attention block is precomputed as XW_Q rather than X. Under mild structural conditions on the surrounding architecture (the LayerNorm obstruction is analysed in Appendix D, and the MLP basis-transfer is investigated empirically in Appendix E.6) this precomputation can be absorbed into the preceding MLP (or, in the first layer, into the embedding), at which point the entire W_Q matrix becomes architecturally redundant. **(ii) Redundancy of learning a linear query and the necessity of nonlinearity.** Every invertible linear W_Q lies on the orbit of I_d, so a learned linear query projection has the same expressive content as the identity, and the parameters spent on learning it are redundant. If one wishes to allocate capacity to the query pathway beyond the identity prior, nonlinearity is imposed by the symmetry argument.

The argument is symmetric in W_Q, W_K, W_V, but we work with W_Q: the KV cache stores precomputed XW_K and XW_V, and Grouped Query Attention (Ainslie et al., 2023) shares W_K, W_V across query groups, so W_Q is the only per-head matrix unaffected by either constraint.

This paper develops both consequences. Section 2 states and proves the reparametrisation lemma and the resulting $W_Q = I_d$ reduction, and treats the multi-layer case under two complementary structural conditions (proofs in Appendix B). Section 3 validates both: GPT-style models with $W_Q = I_d$ match and exceed the full baseline (§3.1), and models with a nonlinear residual query strictly outperform it (§3.2). Section 4 discusses the symmetry-breaking inter-

Code and checkpoints: github.com/MarkoKarbevski/Wqkv_necessity and github.com/MarkoKarbevski/beyond_query_linearity.

pretation, extensions, and limitations.

Related work. Graef (2024) proved that W_Q can be eliminated in skipless transformers via weight reparametrisation; the present work extends this to skip-connected transformers under complementary structural conditions, validates the reduction empirically, and introduces and evaluates the nonlinear query. He & Hofmann (2024) study structural simplifications of transformer blocks, including removal of skip connections and weight matrices; Heimersheim (2024) and Baroni et al. (2025) show that LayerNorm can be removed post-training, providing a path to W_Q elimination in pretrained models. Ji et al. (2025; 2026) show that self-attention is uniquely dependent on its skip connection: removing it causes catastrophic training failure while the MLP degrades more gracefully, which motivates our Theorem B.7 that assumes skip connections exclusively around attention. Kernel-based and linear attention methods (Choromanski et al., 2020) apply nonlinear feature maps to queries but retain a standard linear W_Q projection. Output-side nonlinearities (Qiu et al., 2025) are complementary. Kayyam et al. (2026) study weight tying across the QKV projections (shared key-value, shared query-key, and a single shared projection) and report that on language modelling these reductions are *lossy*: their best variant trades a 50% key-value cache saving for a 3.1% perplexity increase, with downstream metrics degrading correspondingly. Our $W_Q = I_d$ reduction is by contrast exactly lossless, since it removes a provably redundant projection rather than tying two independent ones; the two results are complementary, and the tied-projection setting is the empirical counterpart of the linear $W_K = I_d$ variant we note in Appendix G. Concurrently with the first version of the present work, Wang & Wang (2025) establish the *complete* weight-only symmetry group of canonical multi-head attention as $(\mathrm{GL}(d_k)^h \times \mathrm{GL}(d_v)^h) \rtimes S_h$, proving maximality and extending the analysis to RoPE architectures; see Appendix I for a comparison.

2 The Reparametrisation Symmetry

Notation. Let n be the sequence length, $d = d_{\text{model}}$ the embedding dimension, h the number of heads, and $d_k = d/h$ the head dimension. For each head $i \in \{1, \dots, h\}$ and input $X \in \mathbb{R}^{n\times d}$, define

$$\mathrm{head}_i(X) = \mathrm{softmax}\Big(\frac{(XW_Q^i)(XW_K^i)^\top}{\sqrt{d_k}} + M\Big)(XW_V^i),$$

where $W_Q^i, W_K^i, W_V^i \in \mathbb{R}^{d\times d_k}$ and M is the causal mask. Multi-head attention is then

$$\begin{aligned}&\mathrm{MHA}(X, W_Q, W_K, W_V, W_O)\\ &\quad = \mathrm{Concat}(\mathrm{head}_1(X), \dots, \mathrm{head}_h(X)) \cdot W_O,\end{aligned} \tag{1}$$

where $W_O \in \mathbb{R}^{d\times d}$. We write $W_Q = (W_Q^1 \mid \cdots \mid W_Q^h) \in \mathbb{R}^{d\times d}$, and similarly for W_K, W_V.

Lemma 2.1 (Reparametrisation)**.** *Let $\Theta \in \mathrm{GL}(d)$. Then for all $X \in \mathbb{R}^{n\times d}$,*

$$\begin{aligned}&\mathrm{MHA}(X, W_Q, W_K, W_V, W_O)\\ &\quad = \mathrm{MHA}(X\Theta,\, \Theta^{-1}W_Q,\, \Theta^{-1}W_K,\, \Theta^{-1}W_V,\, W_O).\end{aligned}$$

The proof is elementary but is most transparent in the index-free block notation of Horn et al. (1991); we therefore defer it to Appendix A.2, where the key step, that multi-head attention depends on (X, W_Q, W_K, W_V) only through the three products XW_Q, XW_K, XW_V, is established directly from the definition.

Since the singular matrices form a Lebesgue-null set in $\mathbb{R}^{d\times d}$, $W_Q \in \mathrm{GL}(d)$ almost surely under any initialisation with a density. Setting $\Theta = W_Q$ in Lemma 2.1 then gives

$$\begin{aligned}&\mathrm{MHA}(X, W_Q, W_K, W_V, W_O)\\ &\quad = \mathrm{MHA}(XW_Q,\, I_d,\, W_Q^{-1}W_K,\, W_Q^{-1}W_V,\, W_O)\end{aligned}$$

for all X: the query weight can be replaced by I_d throughout, with $W_K \leftarrow W_Q^{-1}W_K$ and $W_V \leftarrow W_Q^{-1}W_V$.

Remark 2.2 (Parameter savings). Setting $W_Q = I_d$ saves d^2 parameters per layer (25% of standard MHA); realising it in a deep network requires propagating the basis change across layers, the subject of the next paragraph and Appendix B. The lemma is symmetric in W_Q, W_K, W_V, but only one of the three can be fixed to I_d by a single basis change (eliminating two simultaneously would require $W_Q = W_K$, a non-generic condition). With Grouped Query Attention (Ainslie et al., 2023) at $g < h$ K/V groups the fractional saving rises above 25%: 33% at $g = h/2$, 40% at LLaMA-style $g = h/4$, and approaching 50% in the MQA limit ($g = 1$, $h \to \infty$); see Appendix G for the derivation.

Multi-layer propagation. For a transformer block $x \mapsto x + \mathrm{MHA}(x) + \mathrm{MLP}(x + \mathrm{MHA}(x))$, a basis change at one layer modifies the input to the next. Appendix B establishes two sufficient structural conditions under which $W_Q = I_d$ is achievable in *all* layers simultaneously: **(A1)** skip connections only around the attention sublayer (not the MLP), and **(A2)** weight-sharing across layers. The proofs proceed by induction on the layer index; see Theorems B.7 and B.8. Empirically, the reduction holds in standard architectures beyond these conditions (Section 3). Table 1 summarises which architectural components each result covers. The methodology is deliberately three-step: the theorems establish the symmetry *exactly* in a simplified, normalisation-free model (Appendix B); Appendix D establishes that an *approximate* form of the symmetry persists in the full model, the LayerNorm obstruction being mild and the basis change passing through with error vanishing in high dimension (training need not realise this particular approximate symmetry; the theory supplies its existence); and Section 3 tests whether the reduction and its consequences hold *in practice* on the full architecture.

Table 1. Coverage of transformer components in theoretical analysis versus experimental validation. ✓ = feature covered; × = feature simplified or removed; Indep. = result independent of the feature.

Type	Result	Norm.	Skip	Multihead	Multi-layer
Theoretical	Theorem B.2	×	✓	✓	Single layer
	Theorem B.7	×	Around attn. only	✓	✓
	Theorem B.8	×	✓	✓	Shared weights
	Appendix D	✓	Indep.	Indep.	Indep.
Experimental	Section 3	✓	✓	✓	✓

Redundancy of learning a linear query and the necessity of nonlinearity. Lemma 2.1 implies that any learned linear W_Q reduces to I_d via precomputation of XW_Q and absorption of W_Q^{-1} into W_K, W_V, leaving the layer map unchanged. Hence a linear query projection has the same expressive content as the identity: the parameters spent on learning it are redundant. If one wishes to allocate parameters to the query pathway with genuine effect on the function class, the map must be nonlinear. The residual construction $Q(X) = \frac{1}{2}(X + f_\theta(X))$ anchors the nonlinearity to the identity prior while preserving gradient flow (He et al., 2016); the term $f_\theta(X)$ admits no obvious absorption by a $\mathrm{GL}(d)$ transformation.

3 Experiments

Setup. All models are GPT-3-small/NanoGPT-style (Radford et al., 2019; Brown et al., 2020; Karpathy): 12 layers, 12 heads, $d = 768$, GELU activations, LayerNorm, context 1024, GPT-2 BPE tokeniser, no bias parameters; attention is computed with FlashAttention (Dao et al., 2022) throughout. Training uses AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, gradient clipping 1.0, weight decay 0.1, cosine learning-rate schedule (2k warmup, 60k total steps), $\approx$ 490k tokens per step on OpenWebText (Gokaslan & Cohen, 2019) ($\approx$ 29.5B tokens, $\approx 12\times$ Chinchilla-optimal (Hoffmann et al., 2022) for this scale). All configurations share a fixed batch seed and see identical data in identical order. Validation loss is averaged over 2400 sequences ($\approx$ 2.5M tokens) every 1000 steps. Training code and model checkpoints are publicly available; see the title-page footnotes for the repositories. Reported parameter counts include learnable normalisation scales, which amount to $<0.05\%$ of non-embedding parameters and differ negligibly across configurations.

3.1 Symmetry-Reduced Models: $W_Q = I_d$

We train five configurations: the full baseline (124M, 84.95M non-embedding); the symmetry-reduced model ($W_Q = I_d$; 117M, 77.88M non-embedding); two baselines matched to the reduced model's non-embedding count, obtained by shrinking the MLP (hidden $3.5d$) or the model dimension ($d = 744$); and the symmetry-reduced model with the saved parameters reallocated to the MLP (124M, hidden $4.5d$). Results appear in Table 2 and Figure 1.

Three findings emerge. *First*, the 117M symmetry-reduced model matches the full 124M baseline at validation loss 2.944 despite 8% fewer non-embedding parameters: the d^2 parameters of W_Q contribute nothing to the validation log-loss; whether this parity extends to downstream generalisation is an explicit next step (Section 4). *Second*, reallocating saved parameters to the MLP (hidden $4.5d$) achieves the best validation loss across all configurations (2.929), confirming that the MLP is the binding capacity constraint. *Third*, the specificity of the redundancy is confirmed by comparison: standard baselines at the same 78M–80M non-embedding parameter count reach only 2.955–2.956, substantially worse. Reducing the MLP or model dimension costs performance; removing W_Q does not.

The symmetry-reduced architecture benefits from a higher peak learning rate (2.7–3.7× the baseline value of 6×10^{-4}), consistent with the simplified optimisation landscape: attention logits are now linear in W_K rather than bilinear in (W_Q, W_K), so the loss surface admits larger steps. The minimum learning rate is also lower (2×10^{-5} vs. 6×10^{-5}), giving a peak-to-minimum ratio of 80–110× rather than the standard 10×.

3.2 Breaking the Symmetry: Nonlinear Residual Query

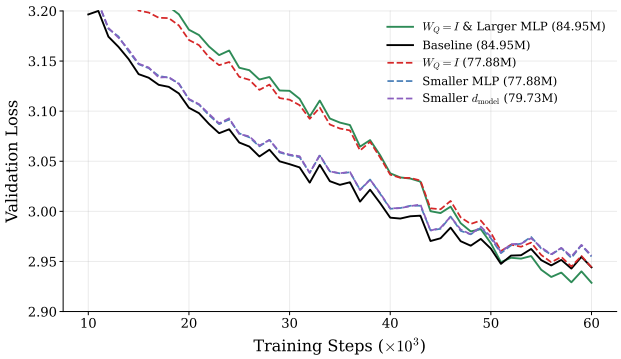
Since any linear W_Q can be absorbed by the reparametrisation symmetry, we replace it with

$$Q(X) = \tfrac{1}{2}\big(X + f_\theta(X)\big), \tag{2}$$

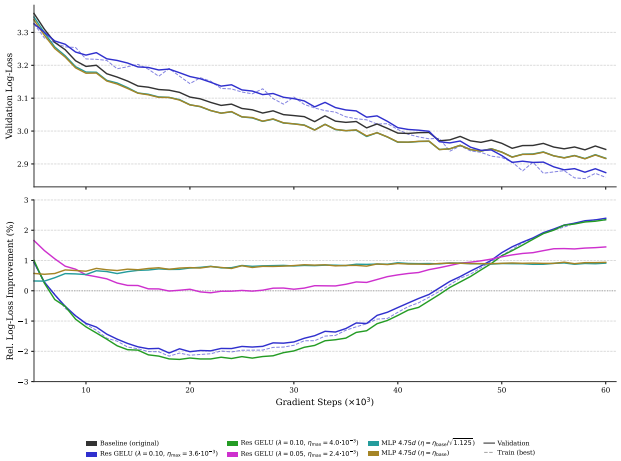
where $f_\theta(X) = \mathrm{LN}(\mathrm{GELU}(\mathrm{RMSNorm}(X)\,W_1)\,W_2)$, $W_1 \in \mathbb{R}^{d\times r}$, $W_2 \in \mathbb{R}^{r\times d}$, $r = d/2$. The parameter count ($d^2 + O(d)$) matches the W_Q it replaces. The identity residual anchors to the $W_Q = I_d$ prior of Section 2 and provides a direct gradient path (He et al., 2016). Keys and values remain standard linear projections. Crucially, the modification is confined to the pre-attention query pathway: the softmax-attention kernel itself is unchanged, so the construction remains compatible with FlashAttention (Dao et al., 2022) and other hardware-aware attention implementations without modification. The added compute is cheap, with the bottleneck nonlinearity matching the standard linear W_Q projection it replaces in both parameter count and FLOPs,

Table 2. All configurations and validation loss at 60k steps ($\approx$29.5B tokens, $\approx$12$\times$ Chinchilla-optimal). †: $W_Q = I_d$, attention scale $1/(2\sqrt{d_k})$. ‡: nonlinear residual query (2); matrix parameters match the 124M baseline exactly, the two normalisation scales of f_θ adding 0.02% (Appendix F). MLP: hidden width of the feed-forward sublayer as a multiple of d (4$\times$ = hidden dimension $4d$).

Configuration	Non-emb.	MLP	Peak LR	Min LR	Val. loss ↓	ΔLL ↑
Baseline (124M)	84.95M	4×	6e-4	6e-5	2.944	ref.
Baseline (117M, MLP 3.5×)	77.88M	3.5×	6e-4	6e-5	2.955	−0.37%
Baseline (118M, $d=744$)	79.73M	4×	6e-4	6e-5	2.956	−0.40%
$W_Q = I_d$ (117M)†	77.88M	4×	1.6e-3	2e-5	**2.944**	+0.00%
$W_Q = I_d$ (124M, 4.5×)†	84.95M	4.5×	2.2e-3	2e-5	**2.929**	+0.52%
MLP 4.75× (+12.5% non-emb.)	95.57M	4.75×	6e-4	6e-5	2.917	+0.94%
Res. NLQ (best)‡	84.97M	4×	3.6e-3	2e-5	**2.874**	**+2.40%**



(a) Symmetry-reduced ($W_Q = I_d$) models vs. baselines.



(b) Nonlinear residual query vs. baseline and the MLP widened from $4d$ to $4.75d$ (+12.5% non-embedding parameters); steps 5k–60k.

Figure 1. Validation loss throughout training. *Left*: the 117M symmetry-reduced model matches the 124M baseline despite 8% fewer non-embedding parameters; reallocation to the MLP (124M, 4.5×) improves further. *Right*: the nonlinear residual query dominates at every step and surpasses a model with 12.5% more non-embedding parameters.

in contrast to alternative routes to comparable expressivity such as widening the MLP. The scope of the claim is structural: the symmetry argument shows that genuine query-side capacity must be *nonlinear*, but it does not single out a particular nonlinearity. Equation (2) is one deliberately simple point in this design space, chosen for exact parameter parity and minimal architectural disruption; we make no claim of optimality. Which nonlinearity is best, and how to compose it, is left to future work.

The best configuration achieves validation loss 2.8736 vs. 2.9441 for the full baseline: a 2.40% reduction in log-loss and a 6.81% reduction in perplexity at no additional parameter cost, comfortably exceeding the gain from widening the MLP hidden layer from $4d$ to $4.75d$, a 12.5% increase in non-embedding parameters (+0.94%). For reference, Biderman et al. (2021) report that replacing learned absolute position embeddings with RoPE (Su et al., 2021) improved log-loss by +1.78% on a 125M GPT-NeoX (Black et al., 2022) model trained for ∼30B tokens on OpenWebText2; our improvement (+2.40%) exceeds this on a comparable setup.

The nonlinear query tolerates peak learning rates up to 3.6×10^{-3} (6$\times$ baseline), consistent with the complete removal of the bilinear $W_Q W_K^\top$ dependence. This stability should not be conflated with QK-normalisation (Henry et al., 2020), which normalises queries and keys *after* their linear projections. QK-norm in fact provides an instructive control: introduced to suppress attention-logit growth at scale (Dehghani et al., 2023), it likewise extends the learning-rate horizon, enabling stable training across three orders of magnitude of learning-rate variation, yet the loss attained at the best-tuned learning rate is essentially unchanged (Wortsman et al., 2024). A wider learning-rate horizon is therefore not, by itself, evidence of added capacity. That the residual nonlinear query both widens the horizon *and* improves the best-tuned log-loss by 2.40% at matched parameter count indicates genuine added expressivity of the query function class, rather than an optimisation side-effect. Consistent with this, the nonlinear query already improves on the baseline at a near-baseline learning rate (peak 8×10^{-4}, min

8×10^{-5}: $+0.44\%$); the full gain is realised at the higher learning rates that the architecture, unlike the baseline, tolerates, and the QK-norm control above shows that such tolerance alone does not produce it (Table 5). Weight decay 0.1 is used throughout; in pre-LayerNorm transformers it acts as an optimisation stabiliser rather than a regulariser (D'Angelo et al., 2024), since it is not applied to the learnable LayerNorm scaling γ.

A practically important observation is that the present implementation is unoptimised: under a restricted compute budget we prioritised training stability and a clean test of the hypothesis over efficiency. It applies two normalisation layers (RMSNorm on input and LayerNorm on output of f_θ) as a defensive choice, to ensure that training instabilities do not confound the architectural ablation; it uses a naive sequential PyTorch forward pass, and has no custom CUDA kernels. Despite this, it trains comparably to, or marginally faster than, the widened MLP (hidden width $4.75d$, $+12.5\%$ non-embedding parameters) in wall-clock time per step, while substantially outperforming it in final validation loss. Several optimisations remain entirely unexplored, and we regard them as important directions for future work: one normalisation layer may suffice; the first bottleneck projection W_1 can be fused with the key and value projections into a single $d \to 2.5d$ matmul (versus the standard fused $d \to 3d$ W_{QKV}), reducing kernel launches and memory traffic at equal total FLOPs, with W_2 applied after the nonlinearity; and the normalisations preceding W_1 can be fused with W_1 into a single normalised-linear kernel. All of these operate on the pre-attention pathway, so the softmax-attention kernel itself (handled by FlashAttention) is unaffected. The current results should therefore be viewed as a conservative lower bound on the performance achievable with the nonlinear query.

4 Discussion and Conclusion

Symmetry-breaking interpretation. The symmetry-reduced baseline ($W_Q = I_d$) selects a canonical representative from the symmetry orbit of the original parametrisation. The nonlinear query (2) *breaks* this symmetry: the map $X \mapsto \frac{1}{2}(X + f_\theta(X))$ does not commute with a $\mathrm{GL}(d)$ basis change, so Lemma 2.1 no longer applies. The query projection thereby acquires genuine expressive content that the reparametrisation of Lemma 2.1 can no longer redistribute to adjacent layers.

MLP Expressivity and Skip Connections. A related question studied in Appendix E concerns when the residual connection around an MLP can be absorbed into a residual-free MLP of the same width. For modern activations, this is either unconditionally impossible (homogeneous activations such as ReLU^2; gated activations such as SwiGLU and GeGLU) or non-generic (ReLU and GELU). This boundary clarifies why saved W_Q parameters are better invested in the MLP than in a linear query projection.

Extensions. Lemma 2.1 is compatible with Rotary Position Embeddings (Su et al., 2021), Grouped Query Attention (Ainslie et al., 2023), Mixture-of-Experts (Shazeer et al., 2017), and QK-normalisation (Henry et al., 2020); details in Appendix G.

Scope. The theoretical results are exact and largely architecture-independent, though variants that factor W_K, W_V through a shared latent projection (e.g. Multi-head Latent Attention) merit a separate analysis. The empirical setup is intentionally conservative (a single architecture, a single seed, matched batches), to ensure the observed gains are not artifacts of tuning. Full-scale validation, multi-seed runs, hyperparameter search, and downstream evaluation are natural next steps, in particular testing whether the $W_Q = I_d$ parity and the nonlinear-query gain persist on downstream tasks; validation log-loss is an appropriate proxy at the present scale (Gadre et al., 2024). Likewise, we defend only the structural claim that query-side capacity requires nonlinearity; optimising the nonlinear query itself (activation, bottleneck width, composition) is beyond the scope of this work and is left to future work and the community. The present results have led us to also explore extending the residual-nonlinearity construction to W_K, W_V, and W_O (see Appendix G). A complementary linear direction follows directly from the reparametrisation of Lemma 2.1: by the same argument that fixes $W_Q = I_d$, one may instead fix the key projection $W_K = I_d$, absorbing it into the query as $\widetilde{W}_Q = W_Q W_K^\top$; and when keys and values are tied ($W_K = W_V$), both can be fixed to I_d simultaneously, with the value projection absorbed into the output map as $\widetilde{W}_O = W_V W_O$. This holds for full-rank attention ($d_k = d$), where setting $W_K = I_d$ is dimensionally admissible. The empirical viability of exactly these tied-projection variants is documented by Kayyam et al. (2026), who find that sharing the key and value projections roughly halves the key-value cache at a modest language-modelling cost (3.1% perplexity); unlike the exact $W_Q = I_d$ reduction, the tied-projection saving is lossy. A natural next step combines the two contributions of this work. We tested the nonlinear query at strict parameter parity with the linear W_Q it replaces; but a lossless linear reduction (either $W_Q = I_d$ or $W_K = I_d$) frees d^2 parameters per layer that need not be returned to the feed-forward sublayer, as in our reallocation experiment, but could instead fund a larger nonlinear query, with a wider bottleneck or greater depth, at no increase in total parameters. Whether a linear reduction on one projection and the nonlinear construction on another compose cleanly, and how best to divide the freed budget between the feed-forward sublayer and the query nonlinearity, is left to future work.

Conclusion. A simplified theoretical analysis and toy-scale empirical evidence support the title's claim: within the linear regime, W_K, W_V alone do suffice, since any one of the three QKV projections can be replaced by the identity at no loss of expressivity. If parameters are to be spent on all three projections, the additional capacity must lie outside the linear regime, and a nonlinear query at parity of parameters, with no search over the nonlinearity itself, already captures substantial gains. Two features of this evidence motivate validation at scale. Parameter-reducing constructions such as Mixture-of-Experts and Grouped-Query Attention typically underperform at toy scale and only catch up at larger sizes; parity for $W_Q = I_d$ already at 124M is therefore atypical. Architectural gains likewise diminish in heavy overtraining ($\approx 12\times$ Chinchilla), yet the residual nonlinear query yields a substantial improvement even in that regime, indicating an expressivity rather than an optimisation effect.

Acknowledgements

We thank the anonymous reviewers of the Weight Space Symmetry workshop for their helpful comments and suggestions. We are grateful to Nils Graef (OpenMachine), Borjan Geshkovski (INRIA/Sorbonne), François Yvon (CNRS/Sorbonne), Yiping Ji (University of Adelaide), and Haris Mandal for valuable discussions on this line of work. The first author also thanks Igor Ševo (HTEC/UniBL) for discussions on the attention mechanism, and Dimitar Peshevski (UKIM) for suggestions on parameter analysis. This work was initiated by the first author while at the HTEC Group, and we thank the team for their support. All experiments presented in this paper and in the accompanying repositories were funded entirely by the first author.

References

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. GQA: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

Ashkboos, S., Croci, M. L., do Nascimento, M. G., Hoefler, T., and Hensman, J. SliceGPT: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations (ICLR)*, 2024.

Baroni, L., Khara, G., Schaeffer, J., Subkhankulov, M., and Heimersheim, S. 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.

Biderman, S., Black, S., Foster, C., Gao, L., Hallahan, E., He, H., Wang, B., and Wang, P. Rotary embeddings: A relative revolution. *EleutherAI Blog*, April 2021. URL https://blog.eleuther.ai/rotary-embeddings/.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., USVSN Sai Prashanth, Purohit, S., Reynolds, L., Tow, J., Wang, B., and Weinbach, S. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of the ACL Workshop on Challenges and Perspectives in Creating Large Language Models (BigScience)*, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

D'Angelo, F., Andriushchenko, M., Varre, A., and Flammarion, N. Why do we need weight decay in modern deep learning? In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 23191–23223. Curran Associates, Inc., 2024. doi: 10.52202/079017-0730. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/29496c942ed6e08ecc469f4521ebfff0-Paper-Conference.pdf.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Dehghani, M. et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning (ICML)*, 2023.

Elhage, N. et al. Toy models of superposition. Transformer Circuits Thread, 2022.

Gadre, S. Y., Smyrnis, G., Shankar, V., Gururangan, S., Wortsman, M., Shao, R., Mercat, J., Fang, A., Li, J., Keh, S., Xin, R., Nezhurina, M., Vasiljevic, I., Jitsev, J., Soldaini, L., Dimakis, A. G., Ilharco, G., Koh, P. W., Song, S., Kollar, T., Carmon, Y., Dave, A., Heckel, R., Muennighoff, N., and Schmidt, L. Language models scale reliably with over-training and on downstream tasks, 2024.

Gemma Team et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Gokaslan, A. and Cohen, V. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

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

He, B. and Hofmann, T. Simplifying transformer blocks. In Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., and Sun, Y. (eds.), *International Conference on Learning Representations*, volume 2024, pp. 8882–8910, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/24fd58f52ff8d0496add8da3991644e9-Paper-Conference.pdf.

He, K., Zhang, X., Ren, S., and Sun, J. 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.

Heimersheim, S. 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.

Henry, A., Dachapally, P. R., Paber, S., and Chen, Y. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35:30016–30030, 2022.

Horn, R. A., Mathias, R., and Nakamura, Y. 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.

Ji, Y., Saratchandran, H., Moghadam, P., and Lucey, S. Always skip attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Ji, Y., Martens, J., Zheng, J., Zhou, Z., Moghadam, P., Zhang, X., Saratchandran, H., and Lucey, S. Cutting the skip: Training residual-free transformers. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.00345.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Karpathy, A. nanoGPT. https://github.com/karpathy/nanoGPT/blob/master/model.py.

Kayyam, A., Gopal, A. M., and Lewis, M. A. Do transformers need three projections? a systematic study of QKV variants. In *International Conference on Machine Learning (ICML)*, volume 306 of *PMLR*, 2026.

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

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Pinkus, A. *Ridge Functions*. Cambridge Tracts in Mathematics. Cambridge University Press, 2015.

Qiu, Z. et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, 2025.

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. 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.

Shazeer, N. GLU variants improve transformer, 2020.

Shazeer, N., Mirhoseini, A., Maziarz, P., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

So, D., Mańke, W., Liu, H., Dai, Z., Shazeer, N., and Le, Q. V. Primer: Searching for efficient transformers for language modeling. *Advances in Neural Information Processing Systems*, 2022.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Team OLMo et al. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Theus, A., Cabodi, A., Anagnostidis, S., Orvieto, A., Singh, S. P., and Boeva, V. Generalized linear mode connectivity for transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 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.

Wang, H. and Wang, K. Complete characterization of gauge symmetries in transformer architectures. In *Symmetry and Geometry in Neural Representations Workshop (NeurReps 2025), NeurIPS*, 2025. URL https://openreview.net/forum?id=KrkbYbK0cH.

Wortsman, M., Liu, P. J., Xiao, L., Everett, K., Alemi, A. A., Adlam, B., Co-Reyes, J. D., Gur, I., Kumar, A., Novak, R., Pennington, J., Sohl-Dickstein, J., Xu, K., Lee, J., Gilmer, J., and Kornblith, S. Small-scale proxies for large-scale transformer training instabilities. In *International Conference on Learning Representations (ICLR)*, 2024.

Xie, Z., Wei, Y., Cao, H., Zhao, C., Deng, C., Li, J., Dai, D., Gao, H., Chang, J., Yu, K., Zhao, L., Zhou, S., Xu, Z., Zhang, Z., Zeng, W., Hu, S., Wang, Y., Yuan, J., Wang, L., and Liang, W. mHC: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*, 2025.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T.-Y. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

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

Zhu, D., Huang, H., Huang, Z., Zeng, Y., Mao, Y., Wu, B., Min, Q., and Zhou, X. Hyper-connections. *arXiv preprint arXiv:2409.19606*, 2024.

Zhu, J., Chen, X., He, K., LeCun, Y., and Liu, Z. 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 Analysis of Multi-Head Attention

A.1 Block Notation for Multi-Head Attention

The index-free block notation introduced here is based on variants of the Block Hadamard Products of Horn et al. (1991). It makes the motivating observation of the paper (that $\mathcal{S}_c$ depends on X only through XW_Q, XW_K, XW_V) immediately transparent from the definition.

Slice operator. For head $i \in \{1, \dots, h\}$, the *slice operator* $P_i : \mathbb{R}^d \to \mathbb{R}^{d_k}$ is the canonical projection onto coordinates $[(i-1)d_k + 1,\ id_k]$. Associativity gives $(XW)^i = XW^i$ for any $W \in \mathbb{R}^{d\times d}$, where superscript i denotes the i-th head block.

Lemma A.1 (Head weights interchange)**.** *For any $X \in \mathbb{R}^{n\times d}$ and $W \in \mathbb{R}^{d\times d}$,*

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

In particular, $(XW_Q)^i = XW_Q^i$ and analogously for W_K, W_V.

Proof. $(XW)^i = (XW)P_i = X(WP_i) = XW^i$ by associativity. □

Block-wise decompositions. We decompose matrices into head-wise blocks of three types.

- $F \in \mathbb{R}^{n\times d}$: column-blocked as $F = (F_1 \mid \cdots \mid F_h)$, each $F_i \in \mathbb{R}^{n\times d_k}$ (e.g. $F = XW_Q$ or $F = XW_V$).
- $G \in \mathbb{R}^{d\times n}$: row-blocked as $G = \left(G_1^\top \mid \cdots \mid G_h^\top\right)^\top$, i.e. G_i are the consecutive d_k-row blocks, each $G_i \in \mathbb{R}^{d_k\times n}$ (e.g. $G = (XW_K)^\top$, so $G_i = (XW_K^i)^\top$).
- $A \in \mathbb{R}^{n\times hn}$: column-blocked as $A = (A_1 \mid \cdots \mid A_h)$, each $A_i \in \mathbb{R}^{n\times n}$ (per-head attention logit or weight matrices).

Note that standard matrix multiplication satisfies $FG = \sum_{i=1}^h F_i G_i$.

Block Hadamard Products. Define two operators:

$$F \boxdot^t G := (F_1G_1 \mid \cdots \mid F_hG_h) \in \mathbb{R}^{n\times hn},$$
$$A \boxdot F := (A_1F_1 \mid \cdots \mid A_hF_h) \in \mathbb{R}^{n\times d}.$$

These are variants of the Hadamard block products of Horn et al. (1991). The *block softmax* applies softmax and masking per head:

$$\mathrm{Softmax}_M^{\boxdot}(A_1 \mid \cdots \mid A_h) := \big(\mathrm{softmax}(M + A_1) \mid \cdots \mid \mathrm{softmax}(M + A_h)\big).$$

Multi-head attention in block notation. Define the self-attention product:

$$\mathcal{S}_c(X, W_Q, W_K, W_V) = \mathrm{Softmax}_M^{\boxdot}\left(\tfrac{1}{\sqrt{d_k}}\,(XW_Q) \boxdot^t (XW_K)^\top\right) \boxdot (XW_V), \tag{3}$$

and $\mathrm{MHA}(X, W_Q, W_K, W_V, W_O) = \mathcal{S}_c(\cdots) \cdot W_O$.

Proposition A.2 (Equivalence with standard definition)**.** *The block-notation $\mathcal{S}_c$ in* (3) *equals* $\mathrm{Concat}(\mathrm{head}_1, \dots, \mathrm{head}_h)$.

Proof. Track the i-th block through each operation. By Lemma A.1, the i-th block of $(XW_Q) \boxdot^t (XW_K)^\top$ is $(XW_Q)^i\big((XW_K)^i\big)^\top = (XW_Q^i)(XW_K^i)^\top$. The block softmax applies $\mathrm{softmax}(M + \cdot)$ to each $n \times n$ block independently, giving $\mathcal{A}_i$ for each i. The block product $\boxdot(XW_V)$ then gives $\mathcal{A}_i(XW_V)^i = \mathcal{A}_i(XW_V^i) = \mathrm{head}_i$. □

A.2 Factorisation of $\mathcal{S}_c$ and Proof of the Reparametrisation Lemma

We now prove Lemma 2.1 via the block notation established above. The argument proceeds in two steps. First, we record the *factorisation observation*: $\mathcal{S}_c$ depends on X only through the three products XW_Q, XW_K, XW_V. Second, we verify directly that the substitution $(X, W_Q, W_K, W_V) \mapsto (X\Theta, \Theta^{-1}W_Q, \Theta^{-1}W_K, \Theta^{-1}W_V)$ preserves these three products simultaneously, and hence preserves $\mathcal{S}_c$. The output projection W_O is unchanged by the substitution, so the conclusion lifts from $\mathcal{S}_c$ to $\mathrm{MHA} = \mathcal{S}_c \cdot W_O$.

Observation A.3 (Factorisation of $\mathcal{S}_c$ through the three products)**.** There exists a function $g : \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}$ such that

$$\mathcal{S}_c(X, W_Q, W_K, W_V) \;=\; g\big(XW_Q,\; XW_K,\; XW_V\big) \quad \text{for all } X \in \mathbb{R}^{n\times d},\; W_Q, W_K, W_V \in \mathbb{R}^{d\times d}.$$

In particular, $\mathcal{S}_c$ does not depend on (X, W_Q, W_K, W_V) individually, only on the three products XW_Q, XW_K, XW_V.

Proof. By Proposition A.2, the block-notation expression (3) is equivalent to the standard concatenation of heads. Setting $A = XW_Q$, $B = XW_K$, $C = XW_V$, equation (3) reads

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

The right-hand side is built from A, B, C, the fixed mask M, the fixed scalar $1/\sqrt{d_k}$, and the block operators $\boxed{\circ}^t$, $\boxed{\circ}$, $\mathrm{Softmax}_M^{\boxed{\circ}}$; it never accesses the factors X or W_Q, W_K, W_V in any way other than through the products A, B, C. Hence g is well-defined as a function of (A, B, C) alone. □

Proof of Lemma 2.1. Fix $\Theta \in \mathrm{GL}(d)$. We compute the three products under the substitution $(X, W_Q, W_K, W_V) \mapsto (X\Theta, \Theta^{-1}W_Q, \Theta^{-1}W_K, \Theta^{-1}W_V)$. For each $P \in \{Q, K, V\}$,

$$(X\Theta)\big(\Theta^{-1}W_P\big) \;=\; X\big(\Theta\,\Theta^{-1}\big)W_P \;=\; X\, I_d\, W_P \;=\; XW_P,$$

using associativity of matrix multiplication and $\Theta\,\Theta^{-1} = I_d$. Therefore the three products $XW_Q,\; XW_K,\; XW_V$ are pointwise unchanged by the substitution. By Observation A.3,

$$\mathcal{S}_c\big(X\Theta,\; \Theta^{-1}W_Q,\; \Theta^{-1}W_K,\; \Theta^{-1}W_V\big) \;=\; g\big(XW_Q,\; XW_K,\; XW_V\big) \;=\; \mathcal{S}_c(X, W_Q, W_K, W_V).$$

Multi-head attention is the post-projection $\mathrm{MHA}(X, W_Q, W_K, W_V, W_O) = \mathcal{S}_c(X, W_Q, W_K, W_V) \cdot W_O$ (cf. the definition immediately below (3)), and the substitution leaves W_O unchanged. Hence

$$\mathrm{MHA}\big(X\Theta,\; \Theta^{-1}W_Q,\; \Theta^{-1}W_K,\; \Theta^{-1}W_V,\; W_O\big) \;=\; \mathrm{MHA}(X, W_Q, W_K, W_V, W_O)$$

for all $X \in \mathbb{R}^{n\times d}$ and all $\Theta \in \mathrm{GL}(d)$, which is the statement of Lemma 2.1. The choice $\Theta = W_Q$ (almost surely well-defined, since $W_Q \in \mathrm{GL}(d)$ for Lebesgue-a.e. W_Q) then gives the $W_Q = I_d$ reduction of Section 2: $\widetilde{W}_K = W_Q^{-1}W_K$, $\widetilde{W}_V = W_Q^{-1}W_V$, and the substituted query is $\Theta^{-1}W_Q = W_Q^{-1}W_Q = I_d$. The roles of Q, K, V are interchangeable: any one of $\Theta = W_Q$, $\Theta = W_K$, $\Theta = W_V$ yields the corresponding identity reduction. □

A.3 Per-Head Block-Diagonal Symmetries

The global $\mathrm{GL}(d)$ action of Lemma 2.1 is not the only weight-space symmetry of multi-head attention. Inside each head there are two further invariances, one acting on the query-key pair and one acting on the value-output pair. Both are well known in the weight-space-symmetry literature (Wang & Wang, 2025) and are complementary to the global action that drives the reductions of Section 2. We record them here for completeness, and to make the joint orbit structure underlying our results explicit.

Proposition A.4 (Query-key block-diagonal invariance)**.** *Let $D = \mathrm{diag}(D_1, \ldots, D_h)$ with each $D_i \in \mathrm{GL}(d_k)$, $d_k = d/h$. Then for all $X \in \mathbb{R}^{n\times d}$,*

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

The orbit of (W_Q, W_K) under this $\mathrm{GL}(d_k)^h$ action has dimension $hd_k^2 = d \cdot d_k$.

Proof. The i-th head's pre-softmax logits are determined by the product $(XW_Q^i)(XW_K^i)^\top$ (cf. Proposition A.2). Under the substitution, head i's logits become

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

using $((D_i^\top)^{-1})^\top = D_i^{-1}$. Every head's logits, and hence the softmax outputs and the $\mathcal{S}_c$ output, are therefore unchanged. The dimension count follows because each D_i ranges over $\mathrm{GL}(d_k)$, contributing d_k^2 free parameters per head. □

A symmetric statement holds on the value-output sector:

Proposition A.5 (Value-output block-diagonal invariance)**.** *Let $M = \mathrm{diag}(M_1, \ldots, M_h)$ with each $M_i \in \mathrm{GL}(d_v)$, $d_v = d/h$. Decompose $W_O \in \mathbb{R}^{d\times d}$ row-wise into blocks $W_O = (W_O^1; \ldots; W_O^h)$ with $W_O^i \in \mathbb{R}^{d_v\times d}$, and write $M^{-1}W_O$ for the matrix whose i-th row block is $M_i^{-1}W_O^i$. Then for all $X \in \mathbb{R}^{n\times d}$,*

$$\mathrm{MHA}\big(X,\, W_Q,\, W_K,\, W_V M,\, M^{-1}W_O\big) \;=\; \mathrm{MHA}(X,\, W_Q,\, W_K,\, W_V,\, W_O).$$

The orbit of (W_V, W_O) under this $\mathrm{GL}(d_v)^h$ action has dimension $hd_v^2 = d \cdot d_v$.

Proof. By Proposition A.2 the i-th head produces $\mathrm{head}_i = \mathcal{A}_i(XW_V^i)$, which is then mapped to the residual stream by the i-th row block of W_O: $\sum_i \mathrm{head}_i \cdot W_O^i$. Under the substitution, head i becomes $\mathcal{A}_i(XW_V^i M_i)$ and its row block of W_O becomes $M_i^{-1}W_O^i$, so the contribution to the output is

$$\mathcal{A}_i(XW_V^i M_i)(M_i^{-1}W_O^i) \;=\; \mathcal{A}_i(XW_V^i)(W_O^i) \;=\; \mathrm{head}_i \cdot W_O^i,$$

which is the original head i contribution. Summing over heads recovers $\mathrm{MHA}(X, W_Q, W_K, W_V, W_O)$. □

Joint group structure. Propositions A.4 and A.5, together with the permutation symmetry S_h which renumbers heads, generate the per-head invariance group of multi-head attention:

$$G_{\mathrm{head}} \;=\; \big(\mathrm{GL}(d_k)^h \times \mathrm{GL}(d_v)^h\big) \rtimes S_h \quad \text{(acting on } (W_Q, W_K, W_V, W_O)\text{)},$$

which is the maximal continuous symmetry of multi-head attention *at the per-head level* (Wang & Wang, 2025). The $\mathrm{GL}(d)$ action of Lemma 2.1 is logically distinct: it acts globally on the input-weight tuple (X, W_Q, W_K, W_V), leaves W_O fixed, and is not contained in G_{head}. The two are complementary, and our parameter-reduction theorems (Appendix B) exploit only the global $\mathrm{GL}(d)$ part; the analogous reduction via G_{head} acting on (W_V, W_O) remains unexplored and is left as future work (see Appendix I).

B Parameter Reduction Theorems

The reparametrisation lemma of the previous section, established via the factorisation observation, suffices to prove our parameter-reduction results. We begin with the single-layer setting, then lift to multi-layer architectures under two complementary structural conditions.

B.1 Single-Layer Query Weight Elimination (Free Lunch)

Throughout this and the next subsection, E and E_P denote the token- and positional-embedding matrices, W_{LM} the LM-head matrix, and $X_0 = E[x] + E_P[0:k]$ the embedded representation of a length-k input x.

The single-layer elimination rests on a structural fact: a block with skip connections around both sublayers commutes with any global change of basis of the residual stream, after an explicit reparametrisation of its weights.

Lemma B.1 (Blocks are conjugate to reduced blocks)**.** *Let B be a transformer block of the form*

$$B(X) = \big(\mathrm{Id} + \mathrm{MLP}\big) \circ \big(\mathrm{Id} + \mathrm{MHA}\big)(X),$$

with MHA *parameters (W_Q, W_K, W_V, W_O), MLP parameters $(W_{\mathrm{up}}, W_{\mathrm{down}})$, and no normalisation. For any $\Theta \in \mathrm{GL}(d)$, define the conjugated block $\widetilde{B}$ with parameters*

$$\widetilde{W}_Q = \Theta^{-1}W_Q, \quad \widetilde{W}_K = \Theta^{-1}W_K, \quad \widetilde{W}_V = \Theta^{-1}W_V, \quad \widetilde{W}_O = W_O\,\Theta,$$

$$\widetilde{W}_{\mathrm{up}} = \Theta^{-1} W_{\mathrm{up}}, \quad \widetilde{W}_{\mathrm{down}} = W_{\mathrm{down}}\,\Theta.$$

Then $\widetilde{B}(X\,\Theta) = B(X)\,\Theta$ for every X, equivalently $B = \Theta^{-1} \circ \widetilde{B} \circ \Theta$ with Θ acting by right-multiplication. Choosing $\Theta = W_Q$ yields $\widetilde{W}_Q = I_d$; since $W_Q \in \mathrm{GL}(d)$ for Lebesgue-a.e. initialisation, this choice is available almost surely.

Proof. By Lemma 2.1, the conjugated attention satisfies

$$\widetilde{\mathrm{MHA}}(X\,\Theta) = \mathcal{S}_c\big(X\Theta,\, \Theta^{-1}W_Q,\, \Theta^{-1}W_K,\, \Theta^{-1}W_V\big)\, W_O\,\Theta = \mathcal{S}_c(X, W_Q, W_K, W_V)\, W_O\,\Theta = \mathrm{MHA}(X)\,\Theta.$$

Hence the attention-plus-skip output is

$$X\,\Theta + \widetilde{\mathrm{MHA}}(X\,\Theta) = (X + \mathrm{MHA}(X))\,\Theta = Z\,\Theta, \qquad Z := X + \mathrm{MHA}(X).$$

The conjugated MLP satisfies $\widetilde{\mathrm{MLP}}(Z\,\Theta) = \sigma(Z\,\Theta\,\Theta^{-1}W_{\mathrm{up}})\,W_{\mathrm{down}}\,\Theta = \mathrm{MLP}(Z)\,\Theta$, so the MLP-plus-skip output is $(Z + \mathrm{MLP}(Z))\,\Theta = B(X)\,\Theta$. □

Theorem B.2 (Single-Layer Query Weight Elimination (Free Lunch))**.** *Consider an L-layer transformer without normalisation, with skip connections around both the attention and the MLP sublayers in every block, untied embedding and LM head, and $W_Q^j \in \mathrm{GL}(d)$ for a chosen layer j. Then there is a reparametrisation of the embedding, the LM head, and the weights of every block that leaves the network's input-output map unchanged and sets $\widetilde{W}_Q^j = I_d$. Concretely, with the single matrix $\Theta = W_Q^j$: every block is conjugated as in Lemma B.1, and $\widetilde{E} = E\,\Theta$, $\widetilde{E}_P = E_P\,\Theta$, $\widetilde{W}_{LM} = \Theta^{-1}W_{LM}$.*

Proof. Let $X_0 = E[x] + E_P[0:k]$ and $X_\ell = B_\ell(X_{\ell-1})$ be the iterates of the original network. In the reparametrised network, $\widetilde{X}_0 = X_0\,\Theta$ by construction, and if $\widetilde{X}_{\ell-1} = X_{\ell-1}\,\Theta$, then Lemma B.1 applied at block ℓ gives $\widetilde{X}_\ell = \widetilde{B}_\ell(X_{\ell-1}\,\Theta) = B_\ell(X_{\ell-1})\,\Theta = X_\ell\,\Theta$. By induction $\widetilde{X}_L = X_L\,\Theta$, and the logits agree: $\widetilde{X}_L\,\widetilde{W}_{LM} = X_L\,\Theta\,\Theta^{-1}W_{LM} = X_L\,W_{LM}$. At layer j, the conjugated query weight is $\Theta^{-1}W_Q^j = I_d$. □

Remark B.3 (Why only a single layer). The same global Θ conjugates every block, but only layer j's query becomes the identity; layer $i \neq j$ retains $\widetilde{W}_Q^i = \Theta^{-1}W_Q^i$. With skip connections around both sublayers, the residual stream carries one basis end to end, so a single global change of variables, and hence a single query elimination, is available. Removing the MLP skip breaks the stream into per-layer segments and permits a fresh basis per layer; this is exactly the mechanism of Theorem B.7. Weight sharing instead makes all layers' queries equal, so the single Θ eliminates all of them at once (Theorem B.8).

Remark B.4. Heimersheim (2024) and Baroni et al. (2025) show that LayerNorm can be removed from pretrained models by fine-tuning. Combining their result with Theorem B.2 provides a post-training path to W_Q elimination.

Remark B.5 (Mildness of the skip-connection assumption). Theorems B.2 and B.7 require a skip connection around the attention sublayer. The assumption is mild: all standard transformer architectures (GPT variants, LLaMA, Qwen, Gemma, OLMo) include such skips by default, and our experiments retain them throughout, so no practically deployed model is excluded. The complementary residual-free regime, while outside the assumption, is not outside the programme: skipless transformers are trainable (Ji et al., 2026), and for them the elimination of W_Q is already established by Graef (2024). Between our theorems and that result, both sides of the assumption are covered. Manifold-Constrained Hyper-Connections (Xie et al., 2025) may further relax the assumption for the original architecture by replacing the identity skip with an invertible mixing matrix; we leave this extension to future work.

B.2 Multi-Layer Query Elimination

Setting (A1): attention-skip-only. An L-layer transformer without normalisation with skip connections only around the attention sublayer:

$$\begin{aligned} Z_\ell &= X_\ell + \mathrm{MHA}_\ell(X_\ell), \\ X_{\ell+1} &= \mathrm{MLP}_\ell(Z_\ell), \quad \ell = 1, \ldots, L, \quad X_1 = X_{\mathrm{in}}. \end{aligned}$$

The block at layer ℓ in setting (A1) is the composition

$$B_\ell(X) \;=\; \mathrm{MLP}_\ell\big(X + \mathrm{MHA}_\ell(X)\big),$$

with $\mathrm{MHA}_\ell(X) = \mathcal{S}_c(X, W_Q^\ell, W_K^\ell, W_V^\ell)\, W_O^\ell$ and $\mathrm{MLP}_\ell(Z) = \sigma(ZW_{\mathrm{up}}^\ell)\, W_{\mathrm{down}}^\ell$ for an element-wise activation σ. The proof of Theorem B.7 rests on a structural identity at the level of the individual block, which we extract as a lemma.

Lemma B.6 (Attention-skip-only block intertwining)**.** *Let B_ℓ be as above with $W_Q^\ell \in \mathrm{GL}(d)$, fix $\Theta_\ell = W_Q^\ell$, and let $\Theta_{\ell+1} \in \mathrm{GL}(d)$ be arbitrary. Define the reduced parameters*

$$\begin{aligned} \widetilde{W}_Q^\ell &= I_d, & \widetilde{W}_O^\ell &= W_O^\ell\, \Theta_\ell, \\ \widetilde{W}_K^\ell &= \Theta_\ell^{-1}\, W_K^\ell, & \widetilde{W}_{\mathrm{up}}^\ell &= \Theta_\ell^{-1}\, W_{\mathrm{up}}^\ell, \\ \widetilde{W}_V^\ell &= \Theta_\ell^{-1}\, W_V^\ell, & \widetilde{W}_{\mathrm{down}}^\ell &= W_{\mathrm{down}}^\ell\, \Theta_{\ell+1}, \end{aligned}$$

and let $\widetilde{B}_\ell$ denote the block built from these reduced parameters. Then for every $X \in \mathbb{R}^{n\times d}$,

$$B_\ell(X)\, \Theta_{\ell+1} \;=\; \widetilde{B}_\ell\big(X\, \Theta_\ell\big). \tag{4}$$

Proof. We trace the block computation step by step under the reduced parameters with input $X\,\Theta_\ell$. By Lemma 2.1 applied with $\Theta = \Theta_\ell$,

$$\begin{aligned} \mathcal{S}_c\big(X\,\Theta_\ell,\, \widetilde{W}_Q^\ell,\, \widetilde{W}_K^\ell,\, \widetilde{W}_V^\ell\big) &= \mathcal{S}_c\big(X\,\Theta_\ell,\, I_d,\, \Theta_\ell^{-1}\, W_K^\ell,\, \Theta_\ell^{-1}\, W_V^\ell\big) \\ &= \mathcal{S}_c\big(X,\, W_Q^\ell,\, W_K^\ell,\, W_V^\ell\big), \end{aligned}$$

hence the reduced multi-head output is

$$\widetilde{\mathrm{MHA}}_\ell(X\,\Theta_\ell) = \mathcal{S}_c(X, W_Q^\ell, W_K^\ell, W_V^\ell)\, W_O^\ell\, \Theta_\ell = \mathrm{MHA}_\ell(X)\, \Theta_\ell.$$

Adding the residual stream gives the reduced MLP input

$$X\,\Theta_\ell + \widetilde{\mathrm{MHA}}_\ell(X\,\Theta_\ell) = X\,\Theta_\ell + \mathrm{MHA}_\ell(X)\,\Theta_\ell = \big(X + \mathrm{MHA}_\ell(X)\big)\,\Theta_\ell = Z_\ell\,\Theta_\ell,$$

where $Z_\ell = X + \mathrm{MHA}_\ell(X)$ is the MLP input in the original block. The reduced MLP then computes

$$\begin{aligned} \widetilde{\mathrm{MLP}}_\ell(Z_\ell\,\Theta_\ell) &= \sigma\big(Z_\ell\,\Theta_\ell\,\widetilde{W}_{\mathrm{up}}^\ell\big)\,\widetilde{W}_{\mathrm{down}}^\ell \\ &= \sigma\big(Z_\ell\,\Theta_\ell\,\Theta_\ell^{-1}\,W_{\mathrm{up}}^\ell\big)\,W_{\mathrm{down}}^\ell\,\Theta_{\ell+1} \\ &= \sigma\big(Z_\ell\,W_{\mathrm{up}}^\ell\big)\,W_{\mathrm{down}}^\ell\,\Theta_{\ell+1} \\ &= \mathrm{MLP}_\ell(Z_\ell)\,\Theta_{\ell+1} \;=\; B_\ell(X)\,\Theta_{\ell+1}, \end{aligned}$$

which is exactly (4). □

Theorem B.7 (Attention-Skip-Only Query Weight Elimination)**.** *Under setting (A1), there exists a simultaneous reparametrisation setting $W_Q^\ell = I_d$ for all ℓ that preserves the network's input-output map. In particular, when the LM head is not tied to the input embedding, the reduced parameters of Lemma B.6 with $\Theta_\ell = W_Q^\ell$, $\Theta_{\ell+1} = W_Q^{\ell+1}$ for $\ell < L$ and $\Theta_{L+1} = I_d$, together with the embedding shift $\widetilde{E} = E\,\Theta_1$, $\widetilde{E}_P = E_P\,\Theta_1$ and unchanged LM head, realise the reduction. When the LM head is tied ($W_{LM} = E^\top$), the same construction works with $\Theta_{L+1} = (\Theta_1^\top)^{-1}$.*

Proof. Let $X_0 = E[x] + E_P[0:k]$ be the embedded input and $X_\ell = B_\ell(X_{\ell-1})$ the iterates of the original network. Choose $\Theta_\ell = W_Q^\ell$ for $1 \le \ell \le L$ and Θ_{L+1} as in the statement; for $1 \le \ell < L$ set $\Theta_{\ell+1} = W_Q^{\ell+1}$. Define the reduced blocks $\widetilde{B}_\ell$ via Lemma B.6 with these parameters, the reduced embeddings $\widetilde{E} = E\,\Theta_1$, $\widetilde{E}_P = E_P\,\Theta_1$, and the reduced LM head

$$\widetilde{W}_{LM} = \begin{cases} W_{LM} & \text{(untied case)}, \\ \widetilde{E}^\top = \Theta_1^\top E^\top & \text{(tied case)}. \end{cases}$$

We claim that for $0 \le \ell \le L$ the iterates $\widetilde{X}_\ell$ of the reduced network satisfy

$$\widetilde{X}_\ell \;=\; X_\ell\,\Theta_{\ell+1}. \tag{5}$$

Base case ($\ell = 0$): $\widetilde{X}_0 = \widetilde{E}[x] + \widetilde{E}_P[0:k] = (E[x] + E_P[0:k])\,\Theta_1 = X_0\,\Theta_1$, so (5) holds with Θ_1 on the right.

Inductive step: assume (5) for $\ell - 1$, i.e. $\widetilde{X}_{\ell-1} = X_{\ell-1}\,\Theta_\ell$. Applying Lemma B.6 at block ℓ,

$$\widetilde{B}_\ell(\widetilde{X}_{\ell-1}) = \widetilde{B}_\ell(X_{\ell-1}\,\Theta_\ell) = B_\ell(X_{\ell-1})\,\Theta_{\ell+1} = X_\ell\,\Theta_{\ell+1},$$

so $\widetilde{X}_\ell = X_\ell\,\Theta_{\ell+1}$, completing the induction.

Output layer. By (5) at $\ell = L$, $\widetilde{X}_L = X_L\,\Theta_{L+1}$. In the untied case $\Theta_{L+1} = I_d$ and $\widetilde{W}_{LM} = W_{LM}$, so the reduced logits are $\widetilde{X}_L\,\widetilde{W}_{LM} = X_L\,W_{LM}$. In the tied case, $\Theta_{L+1} = (\Theta_1^\top)^{-1}$ and $\widetilde{W}_{LM} = \Theta_1^\top E^\top$, giving

$$\widetilde{X}_L\,\widetilde{W}_{LM} = X_L\,(\Theta_1^\top)^{-1}\,\Theta_1^\top\,E^\top = X_L\,E^\top = X_L\,W_{LM}.$$

In both cases the reduced network agrees with the original on every input, while $\widetilde{W}_Q^\ell = I_d$ at every layer. □

We turn to the second multi-layer reduction, where weight tying allows a stronger statement under setting (A2): every block sees both an attention skip and an MLP skip, but the parameters are shared across layers. The key structural fact is Lemma B.1, specialised to $\Theta = W_Q$.

Theorem B.8 (Weight-Shared Query Weight Elimination)**.** *For a transformer with weights tied across all L blocks (each of the form in Lemma B.1), the substitution $\Theta = W_Q$ applied simultaneously at every layer reduces W_Q to I_d while preserving the network's input-output map.*

Proof. Let the shared block be B and apply Lemma B.1 with $\Theta = W_Q$ to obtain a single reduced block $\widetilde{B}$ with $\widetilde{W}_Q = I_d$ such that $B = \Theta^{-1} \circ \widetilde{B} \circ \Theta$. Reparametrise the embeddings and LM head:

$$\widetilde{E} = E\,\Theta, \quad \widetilde{E}_P = E_P\,\Theta, \quad \widetilde{W}_{LM} = \Theta^{-1}\,W_{LM}.$$

Since the same reduced block is used at every layer, the L-fold composition telescopes:

$$B^{\circ L} = (\Theta^{-1} \circ \widetilde{B} \circ \Theta)^{\circ L} = \Theta^{-1} \circ \widetilde{B}^{\circ L} \circ \Theta.$$

At the embedding, $\widetilde{X}_0 = (E[x] + E_P[0:k])\,\Theta = X_0\,\Theta$; after L blocks of the reduced network, $\widetilde{X}_L = \widetilde{B}^{\circ L}(\widetilde{X}_0) = \widetilde{B}^{\circ L}(X_0\,\Theta) = B^{\circ L}(X_0)\,\Theta = X_L\,\Theta$; at the LM head, $\widetilde{X}_L\,\widetilde{W}_{LM} = X_L\,\Theta\,\Theta^{-1}\,W_{LM} = X_L\,W_{LM}$, recovering the original logits. All instances of the shared W_Q in the reduced network equal I_d. □

B.3 Summary: Change of Variables Across the Reductions

Table 3 consolidates the parameter substitutions that realise each of the three reductions established above. Throughout, Θ (without subscript) denotes the global substitution in the single-layer ($\Theta = W_Q^j$ for the chosen layer j, Theorem B.2) and weight-sharing ($\Theta = W_Q$, Theorem B.8) cases, and $\Theta_\ell = W_Q^\ell$, $\Theta_{\ell+1}$ the per-layer substitutions in Theorem B.7. In the single-layer case the same Θ is applied to every block i, and only layer j's query reduces to the identity; the tied LM head is not covered there (Theorem B.2 assumes untied weights).

Table 3. Reduced parameters (denoted $\widetilde{\cdot}$) in each of the three eliminations. In every case $\widetilde{W}_Q = I_d$.

Parameter	Single-Layer (Theorem B.2)	Attention-skip-only (Theorem B.7)	Weight-sharing (Theorem B.8)
W_Q	$\Theta^{-1}W_Q^i$ (I_d at $i{=}j$)	I_d (every ℓ)	I_d
W_K	$\Theta^{-1}W_K^i$	$\Theta_\ell^{-1}W_K^\ell$	$\Theta^{-1}W_K$
W_V	$\Theta^{-1}W_V^i$	$\Theta_\ell^{-1}W_V^\ell$	$\Theta^{-1}W_V$
W_O	$W_O^i\,\Theta$	$W_O^\ell\,\Theta_\ell$	$W_O\,\Theta$
W_{up}	$\Theta^{-1}W_{\text{up}}^i$	$\Theta_\ell^{-1}W_{\text{up}}^\ell$	$\Theta^{-1}W_{\text{up}}$
W_{down}	$W_{\text{down}}^i\,\Theta$	$W_{\text{down}}^\ell\,\Theta_{\ell+1}$	$W_{\text{down}}\,\Theta$
E (embedding)	$E\,\Theta$	$E\,\Theta_1$	$E\,\Theta$
E_P (positional)	$E_P\,\Theta$	$E_P\,\Theta_1$	$E_P\,\Theta$
W_{LM} (untied)	$\Theta^{-1}W_{LM}$	W_{LM}	$\Theta^{-1}W_{LM}$
W_{LM} (tied)	n/a	$\Theta_1^\top E^\top$	$\Theta^{-1}W_{LM}$

In all three cases the reparametrisation is exact: the reduced network agrees with the original network as a function from input tokens to output logits. The single-layer column corresponds to a single block with skip connections around both the attention and the MLP sublayers, with $\Theta = W_Q$ chosen layerwise; the attention-skip-only column tracks the per-layer substitutions required when only the attention sublayer carries a residual connection; and the weight-sharing column collapses the per-layer substitutions to a single $\Theta = W_Q$ shared across the L-fold composition.

C Attention Scaling Correction

With $W_Q = I_d$, the query for token $x \in \mathbb{R}^{1\times d}$ is x itself. At initialisation with $W_K^i \sim \mathcal{N}(0, 1/d)$ elementwise, the logit $(xP_i)\,(xW_K^i)^\top/s$ (with P_i the slice projection of Appendix A.1) has standard deviation of order $\|x\|^2/(s\sqrt{d_k})$. In the standard architecture the logit $(xW_Q)(xW_K)^\top/s$ is of order $\|x\|^2/(s \cdot d_k)$, since two random matrices are contracted; the ratio is $\sqrt{d_k}$. Empirically the observed ratio is $\approx 1.8\times$ (not $\sqrt{d_k} = 8$) due to the joint distribution of x and W_K^i after the embedding layer. Replacing the standard scale $1/\sqrt{d_k}$ with $1/(2\sqrt{d_k})$ restores attention logits to the same order of magnitude as the baseline at initialisation.

D LayerNorm and Change of Basis

Convention. Throughout this appendix, vectors are column vectors and linear maps act by left multiplication, Ax, to align with the standard functional-analysis convention used in the proofs. The main text uses the ML row-vector convention.

Setup. The ε-regularized LayerNorm operator is

$$L_\varepsilon(x) = \frac{x - \mu(x)\mathbf{1}}{\sigma_\varepsilon(x)}, \quad \mu(x) = \tfrac{1}{d}\mathbf{1}^\top x, \quad \sigma_\varepsilon(x) = \sqrt{\tfrac{1}{d}\|x - \mu(x)\mathbf{1}\|^2 + \varepsilon}. \tag{6}$$

The learnable diagonal scaling $D = \mathrm{diag}(\gamma)$ is separated and considered independently; the full LayerNorm reads $D\,L_\varepsilon(x)$. The *zero-mean subspace* is $H = \{z \in \mathbb{R}^d : \mathbf{1}^\top z = 0\}$. The image of L_ε is the open ball $B = \{z \in H : \|z\| < \sqrt{d}\}$, and L_ε admits a unique zero-mean inverse

$$L_\varepsilon^{-1}(z) = \sqrt{\frac{d\varepsilon}{d - \|z\|^2}}\, z, \quad z \in B. \tag{7}$$

Main question. We ask: given a basis change $\Theta \in \mathrm{GL}(d)$ and a diagonal learned scaling D, do there exist a modified MLP MLP' and a diagonal D' such that

$$D'\,L_\varepsilon(x + \mathrm{MLP}'(x)) = \Theta D\,L_\varepsilon(x + \mathrm{MLP}(x)) \quad \forall\, x\,? \tag{8}$$

Equation (8) would mean the modified (LayerNorm + MLP) block carries out the same normalised projection as the original block, up to the basis change Θ.

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

$$L_\varepsilon(f(x)) = DA\,L_\varepsilon(x) \quad \forall\, x \in \mathbb{R}^d. \tag{9}$$

Moreover, $f + h\mathbf{1}$ also satisfies (9) for any function $h : \mathbb{R}^d \to \mathbb{R}$.

Proof. We construct D and f explicitly in three steps.

Step 1 (choose D to preserve H). We need $M = DA$ to satisfy $M(H) \subseteq H$, i.e., $\mathbf{1}^\top(DAz) = 0$ for all $z \in H$. Set $\tilde{v} = (A^\top)^{-1}\mathbf{1}$, $v = \tilde{v}/\|\tilde{v}\|$, and $D_\lambda = \lambda\,\mathrm{diag}(v)$ for $\lambda > 0$. For $z \in H$: $\mathbf{1}^\top D_\lambda Az = \lambda v^\top Az = \lambda(A^\top v)^\top z = (\lambda/\|\tilde{v}\|)\,\mathbf{1}^\top z = 0$.

Choose $\lambda_0 = \|Q^\top \mathrm{diag}(v)AQ\|_2^{-1}$ where $Q \in \mathbb{R}^{d\times(d-1)}$ is an orthonormal basis for H. This ensures $\|M_{0|H}\|_2 = 1$ so that $M_0 = D_{\lambda_0}A$ maps the image ball B into itself.

Step 2 (explicit inverse of L_ε). For $z \in B$, the unique zero-mean pre-image is $L_\varepsilon^{-1}(z)$ given by (7).

Step 3 (define f). Set $f(x) = L_\varepsilon^{-1}(M_0 L_\varepsilon(x))$. Then

$$\begin{aligned} L_\varepsilon(f(x)) &= L_\varepsilon\big(L_\varepsilon^{-1}(M_0 L_\varepsilon(x))\big) \\ &= M_0 L_\varepsilon(x) = D_{\lambda_0} A\,L_\varepsilon(x), \end{aligned}$$

using the fact that L_ε ignores the mean ($f + h\mathbf{1}$ has the same L_ε image as f). Setting $D = D_{\lambda_0}$ completes the construction. □

Theorem D.2 (Basis Transformation Through LayerNorm)**.** *Let $\Theta \in \mathrm{GL}(d)$, $\mathrm{MLP} : \mathbb{R}^d \to \mathbb{R}^d$, and D an invertible diagonal matrix representing the learned LayerNorm scaling. If there exist MLP' and diagonal D' satisfying (8), then D' and MLP' must take the form*

$$\begin{aligned} D' &= \widetilde{D}^{-1}, \\ \mathrm{MLP}'(x) &= L_\varepsilon^{-1}\big(M_0\, L_\varepsilon(x + \mathrm{MLP}(x))\big) - x + h(x)\mathbf{1}, \end{aligned} \tag{10}$$

where $\widetilde{D} = \mathrm{diag}(v)\lambda_0$, $M_0 = \widetilde{D} \cdot \Theta D$, and $h : \mathbb{R}^d \to \mathbb{R}$ is any function. In particular, h contributes a degree of freedom that is passed along the skip connection.

Approximate linearity of f_M in high dimensions. The function $f_M(x) = L_\varepsilon^{-1}(M L_\varepsilon(x))$ is of independent interest as an approximation to the linear map M. Using the SVD $M = U\Sigma V^\top$ and the fact that L_ε is *radial* on H (i.e., $L_\varepsilon(Rx) = RL_\varepsilon(x)$ for orthogonal R preserving H), one shows

$$f_M(x) = U\, f_\Sigma(V^\top x), \quad f_\Sigma(y) = \alpha(y) \cdot \Sigma y, \quad \alpha(y) = \left(1 + \tfrac{\Delta(y)}{d\varepsilon}\right)^{-1/2},$$

where $\Delta(y) = \|y\|^2 - \|\Sigma y\|^2$ is the metric defect. At the origin, $Df_\Sigma(0) = \Sigma$ exactly (the scalings of L_ε and L_ε^{-1} cancel). For a random matrix M with i.i.d. $\mathcal{N}(0, 1/d)$ entries, concentration on the sphere (after clamping singular values to $[0, 1]$, which costs $\|M - M'\|_2 = O(d^{-1/2})$) gives

$$|1 - \alpha(y)| = O\left(\frac{\|y\|^2}{d^{3/2}\varepsilon}\right),$$

so $f_M(x) \approx Mx$ with the approximation improving as $d \to \infty$. This is the mechanism by which the basis change passes through LayerNorm approximately in high-dimensional transformer models.

Remark D.3. RMSNorm (Zhang & Sennrich, 2019), defined as $\mathrm{RMS}_\varepsilon(x) = \gamma \odot x \big/ \sqrt{\frac{1}{d}\|x\|^2 + \varepsilon}$, is covered by the same reasoning: it is a radial map $x \mapsto \phi(\|x\|)\,\mathrm{diag}(\gamma)\,x$, so the orthogonal factors of the SVD pass through exactly and the analysis of f_Σ applies without modification. Dynamic Tanh (Zhu et al., 2025) applies $\tanh$ coordinate-wise with per-channel parameters; it is not radial and the present analysis does not extend to it.

The obstruction from LayerNorm is strictly milder than that of QK-normalisation (Henry et al., 2020): with $W_Q = I_d$, QK-norm normalises the input token representations before computing logits, a standard operation. Numerous state-of-the-art models using QK-norm (Qwen Team, 2025; Gemma Team et al., 2025; Team OLMo et al., 2025) confirm that this is not a practical obstruction.

E When Can Skip Connections Be Absorbed into MLPs?

We study the following question: given a skip-connected MLP $x \mapsto x + \mathrm{MLP}(x)$, does there exist a residual-free MLP $x \mapsto \mathrm{MLP}'(x)$ of the same width computing the same function? The answer depends critically on the activation function.

Setup and notation. We consider single-hidden-layer MLPs

$$\mathrm{MLP}(x) = W_{\mathrm{down}}\, \sigma(W_{\mathrm{up}}\, x), \quad W_{\mathrm{up}} \in \mathbb{R}^{N \times d},\ W_{\mathrm{down}} \in \mathbb{R}^{d \times N},\ N \geq d. \tag{11}$$

The *absorption problem* asks: given $(W_{\mathrm{up}}, W_{\mathrm{down}})$, do there exist $(V_{\mathrm{up}}, V_{\mathrm{down}})$ of the same dimensions such that

$$W_{\mathrm{down}}\, \sigma(W_{\mathrm{up}}\, x) + x = V_{\mathrm{down}}\, \sigma(V_{\mathrm{up}}\, x) \qquad \forall\, x \in \mathbb{R}^d\,? \tag{12}$$

We write $w_i^\top$ for the i-th row of W_{up} and $\hat{w}_i$ for the i-th column of W_{down}.

Gated MLPs.

Definition E.1 (Gated MLP)**.** Let $g : \mathbb{R} \to \mathbb{R}$ be a gate activation. A *gated MLP* maps $\mathbb{R}^d \to \mathbb{R}^d$ by $x \mapsto W_{\mathrm{down}}(g(W_{\mathrm{gate}}\, x) \odot (W_{\mathrm{val}}\, x))$, where $\odot$ denotes the coordinatewise product. SwiGLU ($g = \mathrm{SiLU}$) and GeGLU ($g = \mathrm{GELU}$) are instances of this definition (Shazeer, 2020).

E.1 Reduction to the Identity Skip

Modern architectures replace the identity skip with a learnable mixing matrix M (Zhu et al., 2024; Xie et al., 2025).

Lemma E.2 (Invertible reduction)**.** *If $M \in \mathrm{GL}(d)$, then $Mx + W_{\text{down}}\,\sigma(W_{\text{up}}\,x) = V_{\text{down}}\,\sigma(V_{\text{up}}\,x)$ for all x if and only if $x + (M^{-1}W_{\text{down}})\,\sigma(W_{\text{up}}\,x) = (M^{-1}V_{\text{down}})\,\sigma(V_{\text{up}}\,x)$ for all x.*

Proof. Left-multiply both sides by M^{-1}. □

Hyper-Connection (Zhu et al., 2024) mixing matrices are unconstrained and therefore generically invertible (the set of singular $d \times d$ matrices has Lebesgue measure zero in $\mathbb{R}^{d\times d}$). Manifold-Constrained Hyper-Connections (Xie et al., 2025) use doubly stochastic matrices; since $\det$ is a non-trivial polynomial on the Birkhoff polytope (the identity is doubly stochastic with $\det = 1$), the singular locus has measure zero there too. Hence Lemma E.2 applies to both architectures. All results in this appendix therefore apply to these architectures with the down-projection matrices pre-multiplied by M^{-1}.

E.2 Impossibility for Modern Activations

For the activation functions used in current state-of-the-art models, absorption is unconditionally impossible.

Proposition E.3 (Impossibility: homogeneous activations of degree $k \neq 1$)**.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be homogeneous of degree $k > 0$ with $k \neq 1$, i.e., $\sigma(\lambda z) = \lambda^k \sigma(z)$ for all $\lambda > 0$. Then for any $N, d \geq 1$, any weight matrices, and any $M \in \mathbb{R}^{d\times d}$ with $M \neq 0$, the equation $Mx + W_{\text{down}}\,\sigma(W_{\text{up}}\,x) = V_{\text{down}}\,\sigma(V_{\text{up}}\,x)$ has no solution.*

Proof sketch. Suppose the identity holds for all x. Substituting λx for $\lambda > 0$ and using homogeneity:

$$\lambda Mx + \lambda^k W_{\text{down}}\,\sigma(W_{\text{up}}\,x) = \lambda^k V_{\text{down}}\,\sigma(V_{\text{up}}\,x).$$

When $k > 1$: divide by λ and let $\lambda \to 0$; the MLP terms carry a factor $\lambda^{k-1} \to 0$, leaving $Mx = 0$ for all x, contradicting $M \neq 0$. When $0 < k < 1$: divide by λ^k; as $\lambda \to \infty$ the term $\lambda^{1-k}Mx$ is unbounded unless $Mx = 0$, while both MLP terms stay fixed, again forcing $M = 0$. □

This covers ReLU^2 (used in Primer (So et al., 2022) and Nemotron) and ReGLU (neurons of the form $\mathrm{ReLU}(a^\top x) \cdot (b^\top x)$, homogeneous of degree 2).

Proposition E.4 (Impossibility: gated activations)**.** *Let $g : \mathbb{R} \to \mathbb{R}$ satisfy $g(0) = 0$ and be differentiable at the origin. Then for any $M \neq 0$ and any weight matrices, the equation $Mx + W_{\text{down}}(g(W_{\text{gate}}\,x) \odot W_{\text{val}}\,x) = V_{\text{down}}(g(V_{\text{gate}}\,x) \odot V_{\text{val}}\,x)$ has no solution.*

Proof sketch. Since $g(0) = 0$ and g is differentiable at the origin, $g(z) = g'(0)z + o(z)$ as $z \to 0$, so each gated neuron $g(w^\top x)(v^\top x) = O(\|x\|^2)$. The full gated MLP output is therefore $O(\|x\|^2)$ as $\|x\| \to 0$. Both sides of the absorption equation are $O(\|x\|^2)$, yet $Mx = O(\|x\|)$. Dividing by $\|x\|$ and sending $\|x\| \to 0$ forces $M = 0$. □

Both $g = \mathrm{SiLU}$ (SwiGLU) and $g = \mathrm{GELU}$ (GeGLU) satisfy $g(0) = 0$ and are differentiable at 0. These activations are used in LLaMA (Touvron et al., 2023), Gemma (Gemma Team et al., 2025), Qwen (Qwen Team, 2025), and Mistral (Jiang et al., 2023).

Corollary E.5 (Depth extension)**.** *Let σ be either homogeneous of degree $k > 1$ (applied coordinatewise) or a gated activation satisfying the conditions of Proposition E.4. A depth-L composition of residual blocks $x \mapsto x + \mathrm{MLP}(x)$ cannot equal a depth-L composition of residual-free blocks $x \mapsto \mathrm{MLP}'(x)$ of the same width.*

Proof sketch. In both cases $\mathrm{MLP}(x) = O(\|x\|^2)$ near the origin. For the residual composition, the Jacobian of each block at the origin is I_d, so the composition has Jacobian I_d. For the residual-free composition, each block has Jacobian 0 at the origin, so the composition has Jacobian 0. These cannot agree. □

E.3 Parity Decomposition and the ReLU Case

For ungated activations whose odd part is linear, absorption is *sometimes* possible but fails generically. The key algebraic tool is the following parity decomposition.

Writing $\mathrm{ReLU}(z) = \frac{1}{2}(z + |z|)$:

$$\mathrm{ReLU}(z) = \underbrace{|z|/2}_{\text{even part}} + \underbrace{z/2}_{\text{odd part}} .$$

For the GELU, using $\Phi(-z) = 1 - \Phi(z)$ where Φ is the standard normal CDF, define $E(z) := z(\Phi(z) - \frac{1}{2})$ (even). Then:

$$\mathrm{GELU}(z) = \underbrace{E(z)}_{\text{even part}} + \underbrace{z/2}_{\text{odd part}} .$$

Both activations share the same odd part $z/2$; this produces identical algebraic conditions in Theorems E.6 and E.8.

Theorem E.6 (Identity absorption for ReLU)**.** *Let $d \geq 2$, $N \geq d$. Suppose W_{up} has pairwise non-collinear non-zero rows and W_{down} has non-zero columns. A solution $(V_{\mathrm{up}}, V_{\mathrm{down}})$ to (12) with $\sigma = \mathrm{ReLU}$ exists if and only if there exists $S \subseteq \{1, \ldots, N\}$ with $|S| \geq d$ satisfying*

$$W_{\mathrm{down}}[:, S]\, W_{\mathrm{up}}[S, :] = -I_d. \tag{13}$$

Proof sketch. Sufficiency. Given S with (13), set $D = I_N - 2\Pi_S$ (diagonal, entries ± 1), $V_{\mathrm{up}} = DW_{\mathrm{up}}$, $V_{\mathrm{down}} = W_{\mathrm{down}}$. Since $|Dz| = |z|$ coordinatewise, the even parts agree: $V_{\mathrm{down}}|V_{\mathrm{up}}x| = W_{\mathrm{down}}|W_{\mathrm{up}}x|$. The odd parts give $V_{\mathrm{down}}V_{\mathrm{up}} = W_{\mathrm{down}}(I_N - 2\Pi_S)W_{\mathrm{up}} = W_{\mathrm{down}}W_{\mathrm{up}} - 2I_d$ (using (13)). Combining via $\mathrm{ReLU}(z) = \frac{1}{2}(z + |z|)$: $V_{\mathrm{down}}\mathrm{ReLU}(V_{\mathrm{up}}x) = W_{\mathrm{down}}\mathrm{ReLU}(W_{\mathrm{up}}x) + x$,

Necessity (three steps). Using $\mathrm{ReLU}(y) = \frac{1}{2}(y + |y|)$ and separating even and odd parts uniquely:

$$V_{\mathrm{down}}\, V_{\mathrm{up}} = W_{\mathrm{down}}\, W_{\mathrm{up}} + 2I_d, \tag{R1}$$

$$V_{\mathrm{down}}\, |V_{\mathrm{up}}\, x| = W_{\mathrm{down}}\, |W_{\mathrm{up}}\, x| \quad \forall\, x \in \mathbb{R}^d. \tag{R2}$$

Step 1 (hyperplane alignment). The non-differentiability loci of $x \mapsto W_{\mathrm{down}}|W_{\mathrm{up}}x|$ form the arrangement $\mathcal{H}_W = \bigcup_i \{x : w_i^\top x = 0\}$. Since W_{down} has non-zero columns and W_{up} has pairwise non-collinear rows, all N hyperplanes are distinct and each genuinely contributes a kink. By (R2), $\mathcal{H}_V = \mathcal{H}_W$; the correspondence is a bijection. After relabeling: $V_{\mathrm{up}} = CW_{\mathrm{up}}$, $C = \mathrm{diag}(c_i)$, $c_i \neq 0$.

Step 2 (coefficient matching). Substitute $V_{\mathrm{up}} = CW_{\mathrm{up}}$ into (R2). Using $|CW_{\mathrm{up}}x| = |C||W_{\mathrm{up}}x|$: $V_{\mathrm{down}}|C||W_{\mathrm{up}}x| = W_{\mathrm{down}}|W_{\mathrm{up}}x|$ for all x. Since $d \geq 2$ and W_{up} has pairwise non-collinear rows, the functions $\{|w_i^\top x|\}$ are linearly independent modulo linear functions (distinct kink hyperplanes). Hence $V_{\mathrm{down}} = W_{\mathrm{down}}|C|^{-1}$.

Step 3 (sign extraction). Let $\Sigma = |C|^{-1}C = \mathrm{diag}(\mathrm{sign}\, c_i) \in \{\pm 1\}^{N \times N}$. Substituting into (R1): $W_{\mathrm{down}}(\Sigma - I_N)W_{\mathrm{up}} = 2I_d$. Setting $S = \{i : c_i < 0\}$ gives $\Sigma - I_N = -2\Pi_S$, hence $W_{\mathrm{down}}\Pi_S W_{\mathrm{up}} = -I_d$. Sylvester: $|S| \geq d$. □

E.4 The GELU Case

The GELU proof requires an additional linear independence result.

Lemma E.7 (Independence of scaled analytic functions)**.** *Let $\psi : \mathbb{R} \to \mathbb{R}$ be analytic and not a polynomial, and let $\alpha_1, \ldots, \alpha_m > 0$ be pairwise distinct. If $\sum_k c_k\, \psi(\alpha_k t) = p(t)$ for all t and some polynomial p, then $c_k = 0$ for all k.*

Proof. Write $\psi(t) = \sum_{n \geq 0} a_n t^n$ with $a_n \neq 0$ for infinitely many n. The coefficient of t^n in $\sum_k c_k \psi(\alpha_k t)$ is $a_n \sum_k c_k \alpha_k^n$. For this to vanish for all $n > \deg p$, we need $\sum_k c_k \alpha_k^n = 0$ for infinitely many n. Choosing m such values $n_1 < \cdots < n_m$ gives a linear system whose coefficient matrix $[\alpha_k^{n_j}]_{j,k}$ is a generalised Vandermonde matrix with distinct positive bases, hence invertible. Therefore $c_k = 0$ for all k. □

Theorem E.8 (Identity absorption for GELU)**.** *Under the same hypotheses as Theorem E.6, a solution to (12) with $\sigma = \mathrm{GELU}$ exists if and only if there exists $S \subseteq \{1, \ldots, N\}$ with $|S| \geq d$ satisfying (13).*

Proof sketch. Sufficiency. Identical to the ReLU case: set $D = I_N - 2\Pi_S$, $V_{\text{up}} = DW_{\text{up}}$, $V_{\text{down}} = W_{\text{down}}$. Using $E(-z) = E(z)$ and the parity decomposition: $E(DW_{\text{up}}x) = E(W_{\text{up}}x)$, since D has entries ± 1 and E is even. The odd part contributes $\frac{1}{2}W_{\text{down}}DW_{\text{up}}x = \frac{1}{2}W_{\text{down}}(I_N - 2\Pi_S)W_{\text{up}}x = \frac{1}{2}W_{\text{down}}W_{\text{up}}x + x$, using (13). Hence $V_{\text{down}}\text{GELU}(V_{\text{up}}x) = W_{\text{down}}\text{GELU}(W_{\text{up}}x) + x$,

Necessity (three steps). Applying the parity decomposition to both sides and separating even and odd parts (which are uniquely determined by the identity $\text{GELU}(z) - \text{GELU}(-z) = z$):

$$V_{\text{down}}\, E(V_{\text{up}}\, x) = W_{\text{down}}\, E(W_{\text{up}}\, x), \tag{G1}$$

$$V_{\text{down}}\, V_{\text{up}} = W_{\text{down}}\, W_{\text{up}} + 2I_d. \tag{G2}$$

Steps 1–2 (direction alignment, scaling, and coefficient matching). Project (G1) onto the ℓ-th output coordinate: $\sum_j (V_{\text{down}})_{\ell j}\, E(v_j^\top x) = \sum_i (W_{\text{down}})_{\ell i}\, E(w_i^\top x)$. Group all $2N$ ridge terms by collinearity class of their direction; let $F_k(a_k^\top x)$ be the combined term for each class. Then $\sum_k F_k(a_k^\top x) = 0$ with $\{a_k\}$ pairwise non-collinear. The Pinkus uniqueness theorem (Pinkus, 2015) forces each F_k to be a polynomial; by Lemma E.7, within each class the coefficient of each distinct norm $\|w_i\|$ (i.e., each distinct scaling $\alpha = \|w_i\|$ in $E(\|w_i\| \cdot \hat{w}_i^\top x)$) must vanish separately. From this one deduces: (a) every v_j is collinear with some w_i (else its class is polynomial with vanishing coefficient); (b) for each w_i there is a unique contributing v_j with $v_j = \pm w_i$ and the same column of V_{down} as W_{down}; (c) the correspondence is a bijection. After relabeling: $V_{\text{up}} = CW_{\text{up}}$ with $C = \text{diag}(\pm 1)$ and $V_{\text{down}} = W_{\text{down}}$.

Step 3 (sign extraction, identical to ReLU). Substitute into (G2): $W_{\text{down}}(C - I_N)W_{\text{up}} = 2I_d$. Setting $S = \{i : c_i = -1\}$ gives $C - I_N = -2\Pi_S$, hence (13) with $|S| \geq d$ by Sylvester. □

E.5 Generic Disjointness

Corollary E.9 (Generic disjointness)**.** *Let $d \geq 2$, $N \geq d$, $\sigma \in \{\text{ReLU}, \text{GELU}\}$. For Lebesgue-almost every $(W_{\text{up}}, W_{\text{down}}) \in \mathbb{R}^{N\times d} \times \mathbb{R}^{d\times N}$, no solution to (12) exists.*

Proof. Condition (13) imposes d^2 polynomial constraints on the weights. Since there are finitely many choices of S, the union over all S has Lebesgue measure zero in $\mathbb{R}^{N\times d} \times \mathbb{R}^{d\times N}$. □

Together with Propositions E.3 and E.4, Corollary E.9 shows that the function class of skip-connected MLPs is generically disjoint from that of residual-free MLPs of the same width, for all activation functions encountered in practice. Any method that attempts to represent a residual block as a feedforward block of the same width must either increase the width or accept approximation error.

The following corollary characterises exactly which linear perturbations of the skip can be traded against the MLP. The exceptional family is finite and explicit: sign flips of subsets of hidden units shuttle the corresponding rank-$\geq d$ linear map between the skip and the MLP, and nothing else does.

Corollary E.10 (Linear skips reachable from a given block)**.** *Let $d \geq 2$, $N \geq d$, $\sigma \in \{\text{ReLU}, \text{GELU}\}$, and define $\phi_{W_{\text{up}},W_{\text{down}},Z}(x) := W_{\text{down}}\,\sigma(W_{\text{up}}\,x) + Zx$. Fix $(W_{\text{up}}, W_{\text{down}})$ satisfying the hypotheses of Theorems E.6 and E.8, and let $Z, Z' \in \mathbb{R}^{d\times d}$ with $Z - Z'$ invertible. Then $\phi_{W_{\text{up}},W_{\text{down}},Z} = \phi_{W'_{\text{up}},W'_{\text{down}},Z'}$ for some $(W'_{\text{up}}, W'_{\text{down}})$ if and only if*

$$Z' - Z \;=\; W_{\text{down}}[:, S]\, W_{\text{up}}[S, :] \qquad \textit{for some } S \subseteq \{1, \ldots, N\},\ |S| \geq d.$$

In particular, for Lebesgue-almost every joint tuple $(W_{\text{up}}, W_{\text{down}}, Z, Z')$ the two maps differ.

Proof. Equality of the two maps is equivalent to $W_{\text{down}}\,\sigma(W_{\text{up}}x) + (Z - Z')x = W'_{\text{down}}\,\sigma(W'_{\text{up}}x)$ for all x. Substituting $y = (Z - Z')x$ and setting $B := W_{\text{up}}(Z - Z')^{-1}$, whose rows remain non-zero and pairwise non-collinear (right multiplication by an invertible matrix preserves both), this is the absorption problem (12) for the pair (B, W_{down}) with $W'_{\text{up}}(Z - Z')^{-1}$ in the role of V_{up}. By Theorems E.6 and E.8 it is solvable if and only if $W_{\text{down}}[:, S]\, B[S, :] = -I_d$ for some $|S| \geq d$; right-multiplying by $Z - Z'$ gives the stated condition. When the condition holds, the solution is realised explicitly by the sign-flip construction $W'_{\text{up}} = (I_N - 2\Pi_S)\, W_{\text{up}}$, $W'_{\text{down}} = W_{\text{down}}$. For the final claim: for each of the finitely many S, the condition is a non-trivial polynomial constraint on the joint tuple, so the union over S has Lebesgue measure zero in $(W_{\text{up}}, W_{\text{down}}, Z, Z')$-space. □

We note that the quantifier order matters here: for *fixed* generic weights, the exceptional set of reachable skips $\{ Z + W_{\text{down}} \Pi_S W_{\text{up}} \}_{|S| \geq d}$ is non-empty, so a statement of the form "for almost every $(W_{\text{up}}, W_{\text{down}}, Z)$, no Z' works" would be false; genericity must be imposed jointly on (Z, Z') or the finite family excluded explicitly.

Ungated activations with odd part $z/2$. The identical submatrix condition in Theorems E.6 and E.8 is not a coincidence. Both ReLU and GELU admit parity decompositions with odd part exactly $z/2$, i.e., $\sigma(z) - \sigma(-z) = z$. Any ungated activation satisfying this identity will yield the same condition, provided its even part $E(z) = (\sigma(z) + \sigma(-z))/2$ generates linearly independent ridge functions for non-collinear directions. Ungated SiLU/Swish $\sigma(z) = z/(1 + e^{-z})$ satisfies this; it is rarely used without gating in practice but is covered by the same theorem.

Rank-deficient skip matrices. Lemma E.2 assumes $M \in \text{GL}(d)$. If $\text{rank}(M) = r < d$, write the SVD $M = U \,\text{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0) V^\top$. In rotated coordinates $y = V^\top x$, the problem decomposes: on the r-dimensional subspace corresponding to nonzero singular values, the reduction of Lemma E.2 applies after absorbing the singular values into the down-projection matrices. On the remaining $(d - r)$-dimensional subspace the skip vanishes, and the question reduces to functional equality of two residual-free MLPs on a subspace. For homogeneous activations of degree $k \neq 1$, the degree argument of Proposition E.3 applies for any $x \notin \ker(M)$, so absorption remains impossible whenever $M \neq 0$.

Open problem. Corollary E.5 settles the multi-layer question for homogeneous and gated activations. For ReLU and GELU, whether a depth-L composition of residual blocks can be matched by a depth-L residual-free composition of the same width remains open. The degree argument fails because these activations have a linear term $z/2$ in their parity decomposition, so residual-free blocks at any depth have a nonzero linear part at the origin.

E.6 Supporting Experiment: MLP Basis Transfer Approximation

The theoretical results of this appendix establish exact conditions for skip absorption (impossible for homogeneous/gated activations; non-generic for ReLU/GELU). A natural empirical question is whether gradient descent can nonetheless discover *approximate* solutions for GELU activations in practical dimensions.

Setup. We generate synthetic targets that mirror the basis-change composition arising in the multi-layer propagation argument:

$$Y_{\text{target}}(X) = W_2 \,\text{GELU}(W_1 X) + ZX - X,$$

where $W_1 \in \mathbb{R}^{4h \times h}$, $W_2 \in \mathbb{R}^{h \times 4h}$, and $Z \in \mathbb{R}^{h \times h}$ are randomly initialised. In this subsection h denotes the width of the synthetic problem, matching the figure axes; it is unrelated to the head count of Section 2. We then train a GELU MLP with skip connection of the same shape, $Y_{\text{model}}(X) = W_2' \,\text{GELU}(W_1' X) + X$, to minimise relative ℓ_2 error. We compare against the optimal linear baseline A^* fitted by ridge regression on 500k samples. Experiments are run at $h \in \{256, 512, 768\}$ for 20k AdamW steps (batch size 65,536, peak LR 5×10^{-3}, cosine decay).

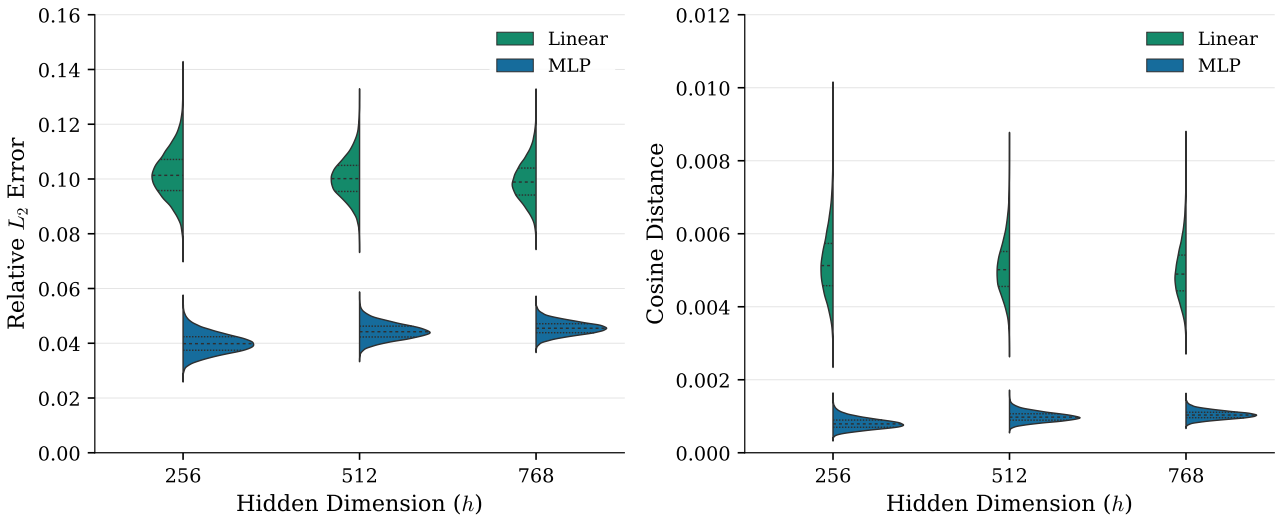


Figure 2. Relative ℓ_2 error and mean cosine distance for approximating a basis-transformed skip-connected MLP. The GELU MLP achieves 4–6% relative error vs. 11–14% for the optimal linear baseline, with cosine similarity ≈ 0.999 across all dimensions.

Results. The trained nonlinear MLP achieves 4–6% relative ℓ_2 error across all three dimensions, outperforming the optimal linear baseline (11–14%) by a factor of roughly $2\times$, with cosine similarity ≈ 0.999 ($\approx 2.5^\circ$ angular error). Performance is consistent across $h \in \{256, 512, 768\}$, suggesting viability at transformer scale. These findings confirm that although exact skip absorption is non-generic, the basis-change composition can be approximated well by gradient descent, bridging the gap between the theoretical guarantees (exact, identity skip) and practice (approximate, basis-transformed skip, full skip).

F Extended Experimental Details

F.1 Full Configuration Tables

Table 4. Complete model configurations and validation loss at 60k steps. MLP: hidden width as a multiple of d. Non-embedding parameter counts follow the conventions of Hoffmann et al. (2022).

Config.	Total	Non-emb.	MLP	Peak LR	Min LR	Val. Loss ↓
Baseline						
(124M)	124.37M	84.95M	$4\times$	6e-4	6e-5	2.944
(117M, MLP $3.5\times$)	117.30M	77.88M	$3.5\times$	6e-4	6e-5	2.955
(118M, $d = 744$)	117.92M	79.73M	$4\times$	6e-4	6e-5	2.956
(140M, MLP $4.75\times$)	134.99M	95.57M	$4.75\times$	6e-4	6e-5	2.917
Symmetry-reduced ($W_Q = I_d$, attention scale $1/(2\sqrt{d_k})$)						
(117M)	117.30M	77.88M	$4\times$	1.6e-3	2e-5	2.944
(124M, MLP $4.5\times$)	124.37M	84.95M	$4.5\times$	2.2e-3	2e-5	**2.929**
Nonlinear residual query $Q(X) = \frac{1}{2}(X + f_\theta(X))$						
WD=0.05, η=2.4e-3	124.39M	84.97M	$4\times$	2.4e-3	3e-5	2.902
WD=0.10, η=4.0e-3	124.39M	84.97M	$4\times$	4.0e-3	1e-5	2.875
WD=0.10, η=3.6e-3	124.39M	84.97M	$4\times$	3.6e-3	2e-5	**2.874**

F.2 Nonlinear Query Architecture

The bottleneck MLP is $f_\theta(X) = \mathrm{LN}(\mathrm{GELU}(\mathrm{RMSNorm}(X)\, W_1)\, W_2)$ with $W_1 \in \mathbb{R}^{d\times r}$, $W_2 \in \mathbb{R}^{r\times d}$, $r = d/2$. For $d = 768$: $r = 384$, giving $2 \times 768 \times 384 = 589{,}824$ matrix parameters, matching $W_Q \in \mathbb{R}^{768\times 768}$ exactly. Normalisation layers add $O(d) < 0.1\%$ overhead. The first bottleneck layer ($d \to r = d/2$) can be fused with the Key and Value projections into a single $d \to 2.5d$ matmul, versus $d \to 3d$ for the standard fused W_{QKV}; the second layer W_2 contributes the remaining $d^2/2$ multiply-accumulates after the nonlinearity, so total FLOPs match the standard pathway and the benefit is fewer kernel launches and less memory traffic.

F.3 Learning Rate Sensitivity

Table 5. Learning rate and weight decay sensitivity for the nonlinear residual query. The baseline diverges before 20k steps at WD = 0.05; the nonlinear variant is more stable throughout.

WD	Peak LR	Min LR	Val. Loss (60k) ↓	ΔLL vs. baseline ↑
0.10	3.6e-3	2e-5	2.8736	+2.40%
0.10	4.0e-3	1e-5	2.8750	+2.35%
0.05	2.4e-3	3e-5	2.9015	+1.45%
0.10	8e-4	8e-5	2.9311	+0.44%
0.10 (baseline)	6e-4	6e-5	2.9441	ref.
0.05 (baseline)	6e-4	6e-5	diverges	n/a

The tolerance of higher learning rates may relate to the scale-invariance induced by normalisation layers (Xiong et al., 2020). This data fills a gap in D'Angelo et al. (2024), whose NanoGPT experiments tested weight decay only in $\{0, 0.1\}$.

F.4 Hyperparameter Sweep

This subsection is illustrative and does not claim completeness. Figure 3 is an earlier snapshot of the sweep, covering roughly 70% of the hyperparameter configurations we ran; the complete set of runs and checkpoints is included in the code release. We include it to expose the sweep conducted and to highlight the unexpectedly high initial gains across configurations, which stabilise mid-training.

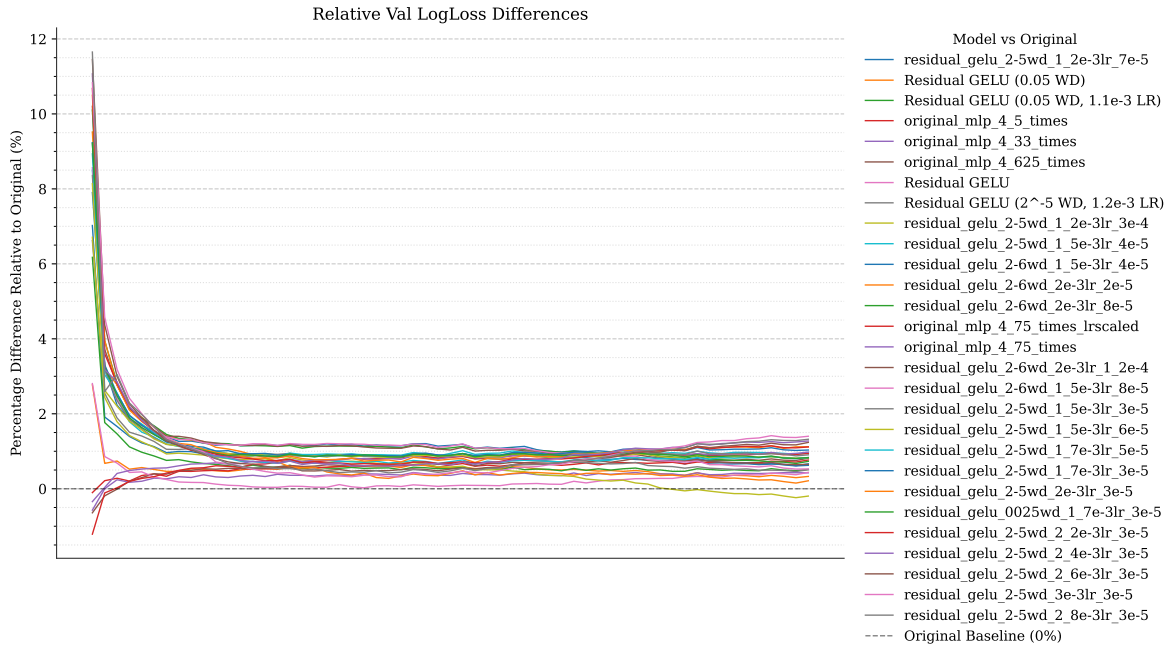
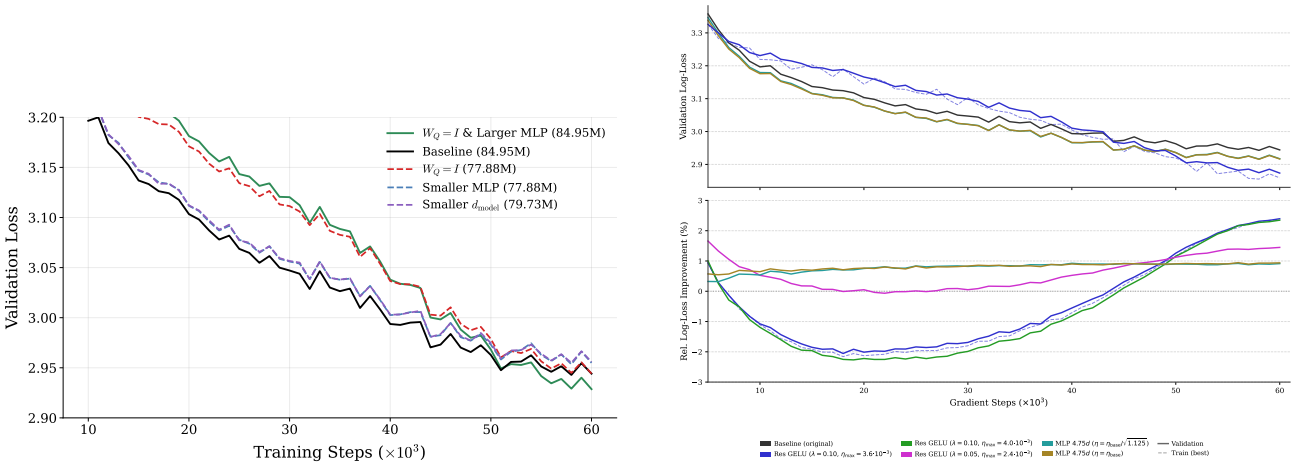


Figure 3. Relative improvement over baseline (steps 1k to 59k). Nonlinear configurations: 84.97M non-embedding parameters; MLP-widened controls: 89.6 to 95.6M.

F.5 Full Training Curves



(a) Symmetry-reduced ($W_Q = I_d$) vs. baselines.

(b) Nonlinear residual query vs. baseline.

Figure 4. Full training dynamics (60k steps). All improvements are consistent throughout training and widen with additional compute.

G Extensions to Modern Architectures

Rotary Position Embeddings. With RoPE (Su et al., 2021), the attention function is

$$g\big(R_m^Q(XW_Q),\ R_n^K(XW_K),\ XW_V,\ W_O\big)\,,$$

where R_m^Q and R_n^K are fixed, position-dependent rotation matrices applied to the queries and keys respectively after the linear projection. The fundamental property underlying Lemma 2.1 is that $\mathcal{S}_c$ is a function of XW_Q, XW_K, XW_V alone.

With RoPE this remains true: define

$$g_1(A, B, C, W_O) = g(R_m^Q A,\ R_n^K B,\ C,\ W_O).$$

Then $\mathcal{S}_c^{\mathrm{RoPE}}(X, W_Q, W_K, W_V, W_O) = g_1(XW_Q, XW_K, XW_V, W_O)$, i.e., the attention output is still a function of XW_Q, XW_K, XW_V alone, now through g_1 rather than g. Lemma 2.1 therefore applies unchanged: any substitution $(X, W_Q, W_K, W_V) \mapsto (X\Theta, \Theta^{-1}W_Q, \Theta^{-1}W_K, \Theta^{-1}W_V)$ preserves XW_Q, XW_K, XW_V and hence $g_1(\cdots)$.

Grouped Query Attention and parameter savings. In GQA (Ainslie et al., 2023) with g key-value groups ($g < h$), W_K and W_V are shared across the h/g query heads per group, so $W_K, W_V \in \mathbb{R}^{d \times g d_k}$ with $d_k = d/h$. The per-layer attention parameter count is then

$$\underbrace{d^2}_{W_Q} + \underbrace{2(g/h)\, d^2}_{W_K,\, W_V} + \underbrace{d^2}_{W_O} \;=\; 2d^2\big(1 + g/h\big).$$

Eliminating W_Q saves exactly d^2 parameters, a fraction $1/(2 + 2g/h)$ of the attention budget; at $g = h/4$ this is 40% of attention parameters per layer. W_Q is the uniquely compatible target under Lemma 2.1: absorbing the $\Theta = W_Q$ basis change into W_K or W_V instead would require separate copies per query head, destroying the GQA compression.

Fixing the key (and value) projection instead. The reduction is not specific to W_Q. Because the attention output depends on X only through XW_Q and XW_K (and through XW_V in the value path), the same reparametrisation that fixes $W_Q = I_d$ applies with the roles of query and key exchanged: setting $\Theta = W_K$ and absorbing it into the query gives $\widetilde{W}_Q = W_Q W_K^\top$ with $W_K = I_d$, leaving the attention logits and hence the full output unchanged. If keys and values are tied ($W_K = W_V$, the shared key-value setting), the single substitution $\Theta = W_K$ fixes both $W_K = I_d$ and $W_V = I_d$ at once, with the value content absorbed into the output projection as $\widetilde{W}_O = W_V W_O$. Two caveats delimit the statement. First, it requires full-rank attention ($d_k = d$): with $W_K = I_d$ the key is X itself, of dimension d, so per-head key dimensions $d_k < d$ leave no identity of the correct shape, the same multi-head obstruction noted above for W_Q in GQA. Second, fixing the query and key projections to a common value is known to symmetrise the attention map and degrade directionality, whereas fixing the key alone (or the tied key-value pair) does not (Kayyam et al., 2026). The empirical counterpart of the tied- projection reduction is studied directly by Kayyam et al. (2026), who report that sharing the key and value projections roughly halves the key-value cache at a modest cost in language-modelling perplexity (3.1%); the saving is lossy, in contrast to the exact $W_Q = I_d$ reduction, because it ties two independent projections rather than removing a redundant one.

Linear Decoupling and Sufficiency of the Skip Connection. The residual nonlinear query (2) introduces a structural decoupling: the query pathway $Q(X) = \frac{1}{2}(X + f_\theta(X))$ is nonlinear, while the key and value pathways $K = XW_K$, $V = XW_V$ remain linear. This separation is principled from the symmetry perspective: the $\mathrm{GL}(d)$ orbit structure means that only the query pathway can break the symmetry productively: adding nonlinearity elsewhere, e.g., in W_K, is absorbed by the same reparametrisation argument via the block-diagonal symmetry of Proposition A.4.

The identity prior in Q is moreover *minimal* within the linear class. By Lemma 2.1, replacing I_d with any other learned linear map (a Hyper-Connection mixing matrix (Zhu et al., 2024), a Manifold-Constrained Hyper-Connection (Xie et al., 2025), or any other learned linear projection) places the parameter triple generically on the orbit of the identity, since the singular linear maps form a measure-zero subset of $\mathbb{R}^{d \times d}$; on this generic set the substitution yields no improvement in expressivity. Productive symmetry breaking therefore requires a nonlinear correction, not a richer linear one.

The skip connection in (2) is in turn necessary to recover full-rank expressivity from the bottleneck. The MLP f_θ factors through a $d/2$-dimensional intermediate representation in our implementation, so

$$\dim_H\big(\mathrm{Im}\, f_\theta\big) \;\le\; d/2$$

in the sense of Hausdorff dimension, for every parameter setting θ: the bottleneck caps the image of the nonlinear branch alone. The identity skip removes this cap, and a single witness suffices: at $f_\theta = 0$ one has $I_d + f_\theta = I_d$, whose image is all of $\mathbb{R}^d$, so

$$\sup_\theta\ \dim_H\big(\mathrm{Im}\,(I_d + f_\theta)\big) \;=\; d,$$

the supremum being attained already at the trivial setting. No genericity or derivative bound is needed; it is enough that the set of θ achieving full dimension is non-empty. The skip is therefore not an optional heuristic but a structural prerequisite for Q to retain the expressivity of the identity baseline: without it the query map is confined to dimension $d/2$, with it the full dimension is already reached at the identity prior.

This decoupling additionally mirrors the outer residual structure of each transformer block: just as $x + \mathrm{MLP}(x)$ couples an identity path with a nonlinear correction, $Q(X) = \frac{1}{2}(X + f_\theta(X))$ couples the identity query prior with a learned nonlinear deviation.

QK-Normalisation. With $W_Q = I_d$, QK-norm (Henry et al., 2020) normalises the input token representations X before computing logits: a standard operation. The obstruction introduced by QK-norm is no stronger than that of standard LayerNorm (see Appendix D). QK-normalisation is moreover orthogonal to the constructions tested in this paper: it extends the stable learning-rate range without improving the best-tuned loss (Wortsman et al., 2024), whereas the nonlinear residual query of Equation (2) changes the query function class and improves the final loss (Section 3).

Mixture of Experts. Lemma 2.1 applies to the attention sublayer independently of the MLP structure: each expert's input projection (in gated variants such as SwiGLU/GeGLU, both the up and gate projections) is conjugated by Θ^{-1}, and its output (down-) projection is right-multiplied by the corresponding basis matrix, exactly as the dense-MLP entries of Table 3. Although attention parameters constitute a small fraction of *total* parameters in a sparse MoE (which is dominated by expert weights), they remain a fully active component of every forward pass, while only a small subset of experts contributes to the MLP computation. The attention-parameter reduction (25% in vanilla MHA, rising to 40% at LLaMA-style GQA and approaching 50% in the MQA limit, both standard in modern MoEs) therefore translates to a non-negligible reduction in *active* per-token compute, in addition to the offloading benefit below. We further conjecture that a Mixture-of-Experts MLP stands to benefit disproportionately from the residual nonlinear query of Equation (2). In a routed MoE layer each token is processed by a small number of experts whose *active* parameter budget must simultaneously accommodate the attention output, the residual stream, and any nonlinear preprocessing the query pathway requires. A nonlinear $Q(X) = \frac{1}{2}(X + f_\theta(X))$ moves the latter into the attention sublayer itself, freeing the limited active expert capacity for the remaining transformations. The same offloading is less consequential for a dense MLP, which is typically wide enough to absorb the additional nonlinear work without competition for capacity.

Extending the construction to W_K, W_V, W_O. The orbit argument is symmetric in W_Q, W_K, W_V, suggesting an analogous residual nonlinear construction $P(X) = \frac{1}{2}(X + g_\theta(X))$ for each. Applying it *simultaneously* to W_Q and W_K is, however, non-trivial: the attention logits become $Q(X)K(X)^\top = (X + f_\theta(X))(X + g_\phi(X))^\top$, whose leading term $XX^\top$ is a parameter-free bias toward self-similarity that dominates at initialisation. A parallel construction at W_V does not share this obstruction, but it interacts with the residual stream geometry; asymmetric designs and Hyper-Connections-style (Zhu et al., 2024; Xie et al., 2025) mixing in the residual stream may resolve this tension. The output projection W_O is in a different position: its elimination is established for the simultaneously skipless and LayerNorm-free regime by Graef (2024), suggesting that the same residual nonlinear construction may apply once those architectural simplifications are in place. Empirical validation of all of these extensions is left to future work.

Arbitrary attention masks. The mask M enters Lemma 2.1 as a fixed term, unaffected by the substitution; the symmetry therefore holds for any choice of mask, including the bidirectional ($M = 0$) encoder-only setting and prefix-LM masks. Relative parameter savings tend to be larger in encoder-only models, which typically use smaller MLP hidden dimensions.

H On the Limitations of the Linear $\mathbb{R}^d \to \mathbb{R}^{3d}$ QKV Pathway

The empirical gain from replacing the linear W_Q with the nonlinear residual f_θ (2.40% lower validation log-loss at 124M parameters) is materially larger than what neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) predict for parameter increases of comparable size (our own control measures 0.94% for an MLP widened from $4d$ to $4.75d$, a 12.5% non-embedding parameter increase; Chinchilla-style fits in this regime extrapolate similar marginal returns), and exceeds what reasonable seed-level stochasticity could account for. The algebraic redundancy established in the present work and by Graef (2024) explains why a linear W_Q is wasted, but does not by itself explain why a parameter-equivalent nonlinear replacement should outperform a similarly-sized linear architecture by such a margin.

This appendix sketches two complementary, preliminary perspectives that attempt to formalise this gap. The first is informal and rests on a Linear-Representation-Hypothesis-style assumption (Park et al., 2024; Elhage et al., 2022); the second is a formal observation about the affine geometry of the QKV map and requires no such assumption. Both are speculative and deserve more careful development than we give them here; in particular, neither is sufficient on its own to predict the magnitude of the observed gain, only to identify structural reasons it should be nonzero. We include them to make the conjecture concrete and to provide a starting point for follow-up investigation.

H.1 The Concept-Space Picture (informal)

Suppose, in the spirit of the Linear Representation Hypothesis, that the immediate upstream MLP+Id block writes features into the residual stream along learned concept directions $(c_i)_{i\in I}$. The input to the attention's QKV projection can then be written as $y = \sum_{i\in I} a_i c_i$, with concept coefficients (a_i) that depend on the token. Under this view, the linear $W_{QKV} = [W_Q; W_K; W_V]$ pathway faces two related structural difficulties.

Roadblock 1: one y must serve three pathways. By linearity, $W_P y = \sum_i a_i (W_P c_i)$ for each $P \in \{Q, K, V\}$. The token-dependent coefficients a_i are identical across the three pathways; each W_P can only re-express the concept basis as $(W_P c_i)$ but cannot reweight a_i pathway-by-pathway. The upstream MLP+Id is therefore solving an encoding problem: pack into d dimensions a concept activation pattern that simultaneously serves three distinct downstream uses (querying, being queried, contributing value).

Roadblock 2: no nonlinear combination downstream. Any nonlinear combination of concept activations that any of the three pathways might need must be precomputed by the MLP+Id and encoded directly in y, because the path from y through W_{QKV} to the attention softmax is purely linear. If the query pathway needs to attend conditionally on a conjunction of concepts (e.g., "anger and question"), that conjunction must be allocated its own concept dimension in y at the MLP+Id stage. There is no architectural component between the MLP+Id and the attention softmax that can synthesise such conjunctions from constituent activations.

Together, the two roadblocks suggest that pathway-specific nonlinearity placed between the MLP+Id and attention addresses both at once: it allows pathway-specific concept reweighting (Roadblock 1) and pathway-specific nonlinear combination (Roadblock 2). The nonlinear residual query proposed in this work is one realisation. A natural alternative reuses the upstream MLP+Id's already-computed nonlinear hidden state $\mathrm{GELU}(W_{\mathrm{up}}\cdot) \in \mathbb{R}^{4d}$: it contains $4d$ nonlinearly-computed features of its input, of which only a single d-dimensional linear combination reaches the next attention via W_{down}. Routing this hidden state (or a d-dimensional slice of it) directly into the next attention's query pathway, concretely $Q(X) = (X + u^{(\ell-1)} W'_Q)/2$ where $u^{(\ell-1)} = \mathrm{GELU}(W_{\mathrm{up}}^{(\ell-1)}\cdot)$ is the previous block's MLP hidden state, gives the query access to a $4d$-dimensional nonlinear basis that the K and V pathways do not see (addressing Roadblock 1) and that is already nonlinearly composed (addressing Roadblock 2), at the cost of one additional linear projection and no recomputed nonlinearity. The proposed f_θ is the more conservative, intra-block realisation of the same architectural intuition; the MLP-reuse variant is left to future work.

A more aggressive refinement of the MLP-reuse design widens the MLP output to $\mathbb{R}^{d+d_Q+d_K+d_V+d_{Q,K}+d_{Q,V}+d_{K,V}}$, where the standard d-dimensional component continues to feed the residual stream and the remaining subspaces are routed to specific pathways of the next attention block: per-pathway dimensions d_Q, d_K, d_V are visible only to a single pathway, and pair-shared dimensions $d_{Q,K}, d_{Q,V}, d_{K,V}$ are visible to a pathway pair (so the attention-logit pathway sees $d + d_{Q,K}$ shared dimensions plus its individual d_Q or d_K, and so on). This is the most explicit architectural realisation of pathway-specific concept access from Roadblock 1: each pathway is allocated its own designated concept subspace by construction, with no shared W_Q, W_K, W_V projection forcing the same basis on all three. The allocation of dimensions across the seven subspaces is a design choice that requires empirical investigation, as does the integration with attention-head structure; we leave both the architecture and the training dynamics to future work.

This picture rests on a substantive assumption: that something like the Linear Representation Hypothesis holds at the relevant layer and architectural depth, and that the MLP+Id is the operative concept-writing component. Both claims have empirical support from the mechanistic interpretability literature but neither is established for the specific architecture studied here. Whether trained f_θ in fact performs pathway-specific concept reweighting (as opposed to other forms of useful computation) is an empirical question we leave open.

Geometric restatement. The two roadblocks above admit a coordinate-free restatement. The upstream MLP+Id, however expressive its nonlinear computation, can only produce content that the downstream linear W_{QKV} pathway then projects into a d-dimensional linear subspace of the $3d$-dimensional ambient space inhabited by (Q, K, V). The MLP+Id is therefore bottlenecked not only by its own output width d but, downstream, by the fact that the linear W_{QKV} map cannot reach more than a d-dimensional slice of the available $3d$ ambient capacity; the other $2d$ ambient directions are unreachable by any choice of linear projections. This restriction is materially stronger than "d-dimensional image in a $3d$-dimensional ambient" might suggest. A d-dimensional linear subspace is flat: every output point lies in the same fixed d-dimensional slice, regardless of input. A d-dimensional curved submanifold (or even a topologically d-dimensional set produced by a nonlinear map) can wind through $\mathbb{R}^{3d}$ with its affine hull occupying up to the full $3d$ ambient capacity; a 1-dimensional helix

in $\mathbb{R}^3$ is the canonical illustration, locally 1-dimensional yet globally spanning all three ambient axes. The next subsection makes this geometric constraint precise without invoking the LRH-style assumption used above.

H.2 The Affine Hull Bound (formal)

A complementary observation concerns the geometric extent of the QKV pathway's image in $\mathbb{R}^{3d}$, without reference to learned features or concept bases.

Definition (affine hull dimension). For a set $S \subseteq \mathbb{R}^N$, the affine hull dimension is

$$\dim \operatorname{aff}(S) := \dim \operatorname{span}\{\, s - s_0 : s \in S \,\} \quad \text{(for any fixed } s_0 \in S\text{)},$$

i.e., the dimension of the smallest affine subspace containing S. This is distinct from the topological dimension of S as a manifold and from its Hausdorff dimension (the notion used in the skip-connection argument of Appendix G). A 1-dimensional helix in $\mathbb{R}^3$ has topological and Hausdorff dimension 1 but affine hull dimension 3: locally it has one degree of freedom, but globally it spans the full ambient space. A space-filling curve (Peano, Hilbert) shows the extreme: 1-dimensional continuous parametrisation, affine hull equal to the entire ambient $\mathbb{R}^k$.

Setup. Consider a generalised QKV-style map

$$\Phi : \mathbb{R}^d \to \mathbb{R}^{3d}, \qquad \Phi(x) = (f(x),\, Px,\, Qx),$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ is possibly nonlinear and $P, Q : \mathbb{R}^d \to \mathbb{R}^d$ are linear. The standard fully linear QKV pathway corresponds to f also linear; the architecture proposed in this work corresponds to f nonlinear with P, Q retained as the linear key and value projections.

Proposition H.1 (Affine hull bound)**.** *Let $S = \Phi(\mathbb{R}^d)$. Then*

$$\dim \operatorname{aff}(S) \;\le\; \dim \operatorname{span}\big(\operatorname{im}(f) - f(0)\big) + \operatorname{rank}([P; Q]) \;\le\; 2d,$$

where $[P; Q] \in \mathbb{R}^{2d \times d}$ is the vertical stacking of P and Q.

Proof. Let $W = \operatorname{span}\{\Phi(x) - \Phi(0) : x \in \mathbb{R}^d\}$, so $\dim \operatorname{aff}(S) = \dim W$. Decompose $\mathbb{R}^{3d} = \mathbb{R}^d \oplus \mathbb{R}^{2d}$ into the f-coordinates and the (P, Q)-coordinates, with projections π_1, π_2. From $\Phi(x) - \Phi(0) = (f(x) - f(0),\, Px,\, Qx)$,

$$\pi_1(W) \subseteq \operatorname{span}\big(\operatorname{im}(f) - f(0)\big), \qquad \pi_2(W) = \operatorname{im}([P; Q]).$$

Applying the sub-additivity bound $\dim W \le \dim \pi_1(W) + \dim \pi_2(W)$ (rank-nullity for $\pi_2|_W$, with $\ker(\pi_2|_W) \subseteq \pi_1(W) \oplus 0$) gives the claim. □

The bound treats π_1 and π_2 as independent, while both depend on the same x, and is therefore loose when this coupling is strong. The extreme case is linear f: then Φ is itself a single linear map $\mathbb{R}^d \to \mathbb{R}^{3d}$ of rank at most d, so $\dim \operatorname{aff}(S) \le d$, half of what the sub-additivity bound asserts. The bound is approached only when f is nonlinear in a way that lets $\pi_1(W)$ contribute directions decoupled from $\pi_2(W)$; a tight characterisation of this coupling is left to future work.

Three architectural regimes. Specialising to the three placements of nonlinearity in the QKV pathway: for a fully linear pathway (f also linear), the entire map Φ is linear and $\dim \operatorname{aff}(S) \le d$; the image is contained in a d-dimensional linear subspace of $\mathbb{R}^{3d}$, leaving $2d$ ambient dimensions unreached. With one nonlinear pathway (f nonlinear, P and Q linear, the architecture in this work), $\dim \operatorname{aff}(S) \le 2d$, with equality approached for sufficiently nonlinear f. With all three nonlinear, $\dim \operatorname{aff}(S) \le 3d$ by an analogous argument, with similar conditions for tightness. The qualitative point is that moving from the all-linear case to the one-nonlinear case can unlock up to d additional ambient dimensions of geometric extent, conditional on f being substantively nonlinear; the marginal return on additional nonlinear pathways is similarly conditional and not automatic.

Interpretation and limitations. The upper bound rests on linear-algebraic and finite-difference facts and requires no assumption about learned features or the structure of the residual stream. However, it is an upper bound, not an achievability statement, and a precise characterisation of when nonlinearity actually unlocks additional ambient dimensions (in particular, a clean quantitative measure of "substantive nonlinearity" relating to the variation of f across the input distribution) is left to future work. Whether trained f_θ actually exploits this geometric headroom, and whether the geometric gain corresponds to the empirical performance improvement we observe, are open empirical questions.

I Connection to Weight-Space Symmetry Literature

The $\mathrm{GL}(d)$ symmetry of Lemma 2.1 is an instance of the broader class of weight-space symmetries in neural networks. Permutation symmetries in MLPs arise from the discrete group S_N acting on hidden units. Scaling symmetries from normalisation layers arise from the positive reals acting on preactivations. The per-head block-diagonal symmetry (Proposition A.4) is a product of h copies of $\mathrm{GL}(d_k)$. Our result identifies a global $\mathrm{GL}(d)$ action that acts jointly on the residual stream and the attention weights (X, W_Q, W_K, W_V), which is the primary object of study in this paper.

Relation to Wang & Wang (2025). Published concurrently with the first version of the present work, Wang & Wang (2025) prove that the maximal *weight-only* symmetry group of canonical multi-head attention is

$$G_{\max} = \left(\mathrm{GL}(d_k)^h \times \mathrm{GL}(d_v)^h\right) \rtimes S_h, \tag{14}$$

where the $\mathrm{GL}(d_k)^h$ factor acts per-head on the query-key sector as $(W_Q^i, W_K^i) \mapsto (W_Q^i A_i, W_K^i (A_i^{-1})^\top)$ and $\mathrm{GL}(d_v)^h$ acts on the value-output sector as $(W_V^i, W_{O,i}) \mapsto (W_V^i C_i, C_i^{-1} W_{O,i})$. The S_h factor permutes heads.

Two structural observations connect their result to ours. *First*, the $\mathrm{GL}(d_k)^h$ factor in $G_{\max}$ is precisely the per-head block-diagonal symmetry of Proposition A.4: their right-action $W_Q^i \mapsto W_Q^i A_i$ preserves $W_Q^i (W_K^i)^\top$ when combined with the compensating key transform, in the same way that our left-action $W_Q^i \mapsto \Theta^{-1} W_Q^i$ preserves the attention output. *Second*, the symmetry groups act on different objects. The group $G_{\max}$ acts on the weight matrices *at fixed input* X: it is the maximal group of parameter transformations that preserve the function for all X. Our $\mathrm{GL}(d)$ symmetry (Lemma 2.1) acts *jointly* on (X, W_Q, W_K, W_V): it is a symmetry of the attention *map composition* rather than of the weights alone. As a consequence, our $\mathrm{GL}(d)$ transformation is not a subgroup of $G_{\max}$ (which fixes X), but is instead a symmetry of the input-parameter pair. This is why Lemma 2.1 enables the architectural reduction $W_Q = I_d$: by choosing $\Theta = W_Q$ and simultaneously replacing X with XW_Q, we move to an equivalent input representation in which the query projection is the identity.

For RoPE architectures, Wang & Wang (2025) show that the query-key sector of $G_{\max}$ reduces from $\mathrm{GL}(d_k)^h$ to $(\mathcal{C}_{\mathrm{RoPE}})^h \cong (\mathrm{GL}(1, \mathbb{C}))^{d_k/2}$, where $\mathcal{C}_{\mathrm{RoPE}}$ is the commutant of the position-dependent rotations (see Appendix G for how Lemma 2.1 applies to RoPE via the g_1 reformulation).

Relation to symmetry-based model alignment. Theus et al. (2025) exploit the same residual-stream change of basis for a different purpose: aligning independently trained transformers to reveal linear mode connectivity. Their framework restricts the residual-stream transformation to the orthogonal group, because RMSNorm is exactly invariant under orthogonal maps and LayerNorm reduces to RMSNorm once its mean-centring and diagonal scale are absorbed into the adjacent linear layers (Ashkboos et al., 2024); the per-block conjugation they apply is the orthogonal specialisation of Lemma B.1. Our Appendix D is complementary: in its radial SVD analysis, the orthogonal factors of a general $\Theta \in \mathrm{GL}(d)$ pass through the normalisation exactly, which is precisely the mechanism their alignment exploits, while the singular-value part contributes the obstruction we quantify, with error vanishing in high dimension. The two uses of the symmetry also differ in aim: alignment searches for a transformation relating two trained models, whereas our reductions fix the transformation $\Theta = W_Q$ that eliminates a parameter block within a single model.

$W_Q = I_d$ breaks both symmetries. The reduction $W_Q = I_d$ of Section 2 simultaneously trivialises two of the symmetries above. The global $\mathrm{GL}(d)$ left-action of Lemma 2.1 preserves $W_Q = I_d$ only when $\Theta = I_d$, since $\Theta^{-1} I_d = \Theta^{-1}$. The per-head right-action $W_Q^i \mapsto W_Q^i A_i$ from $\mathrm{GL}(d_k)^h$ similarly preserves the i-th column slab of I_d only when $A_i = I_{d_k}$. Hence $W_Q = I_d$ is a fixed point of trivial stabiliser under both groups, and is in this sense a fully symmetry-reduced representative on the QK side.

Future work: the (W_V, W_O) reduction and a W_K/W_O duality. The value-output symmetry of Proposition A.5 identifies a $\mathrm{GL}(d_v)^h$ redundancy in (W_V, W_O) that mirrors the query-key reduction exploited in this paper; carrying out the corresponding elimination is a concrete direction for future work. The maximal symmetry $G_{\max}$ of (14) treats the QK and VO sectors as independent products, but the matrices W_K and W_O play structurally analogous roles: both are $d \times d$ projections mediating between the residual stream and per-head subspaces, on the input and output sides of the attention computation. Whether this analogy admits a formal symmetry connecting the two sectors, and whether the residual nonlinear construction of Equation (2) admits W_O-side counterparts in the spirit of Graef (2024), remains open.