WHAT CAN BE IDENTIFIED IN A TRANSFORMER?

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic interpretability names parts of a network (heads, neurons, residual directions), yet many weight settings compute the same function. We ask which of these parts the function determines, and obtain three results. First, a multi-head attention layer is identified exactly: its function determines precisely the measure that assigns to each query–key form the summed value–output maps of the heads using it, already from inputs of length three, so head permutations and per-head gauges are its only symmetries on generic weights. Second, the symmetries of deployed transformers depend on their architecture: rotary embeddings, QK-normalization and attention sinks constrain the gauges, and normalization placement decides whether the residual stream has a privileged basis (none in pre-norm models, only signed permutations when sublayer outputs are normalized). We verify every prediction, with a negative control, on nine checkpoints of seven families up to 32B parameters in float64. Third, these symmetries sort interpretability claims: claims about a checkpoint must be equivariant, claims about its function invariant, and claims about what training produces must replicate across runs. Across independently trained seeds of Pythia and GPT-2, head indices carry no information about IOI or induction attributions, whereas heads matched by symmetry-invariant signatures, and label-invariant statistics, replicate.

1 INTRODUCTION

Mechanistic interpretability explains a network by naming its parts: "head 9.6 moves the indirect object" (Wang et al., 2023), "this head is an induction head" (Olsson et al., 2022), "residual dimension 1415 carries a massive activation" (Sun et al., 2024), "this direction encodes sentiment" (Park et al., 2024). Each of these objects is defined through a particular set of weights. But many weight settings compute the same function, and a statement about a head, a neuron or a basis direction can describe the chosen parametrization rather than the computation: a change of GPT-2's weights that leaves every output unchanged reverses the ranking of its 144 heads by the norm of their output weights (Section 4.4). This paper asks, for transformers as built and deployed, *which of these objects the function determines*, and turns the answer into a checkable criterion for interpretability claims.

Three kinds of claims. A claim about *a checkpoint* ("in this file, head 9.6 moves names") only needs to be *equivariant*: relabelling the heads relabels the claim. A claim about *the function* must be *invariant* under every weight change that preserves the function. A claim about *what training produces* ("language models learn induction heads") must survive independent training runs, whose labels and bases are arbitrary relative to one another. The symmetry group of the architecture decides which statements fall in which class (Section 3).

Contributions.

- **Complete identifiability of an attention layer** (Section 4.1, Figure 1a). The layer's function determines exactly one object: the finitely supported measure that assigns to each query–key form the sum of the value–output maps of the heads that use it (Theorem 4.1); inputs of length three suffice, and length two does not. Hence head permutation together with the per-head query–key and value–output gauges is the *complete* symmetry group on an open dense set of weights, and we describe every fibre off it (Corollary 4.2).
- **The architecture decides the group** (Sections 4.1–4.3). RoPE restricts the query–key gauge to scaled rotations of frequency pairs; grouped-query attention ties gauges within a group; QK-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

normalization moves the gauge into the normalization gains; RoPE and attention sinks make the key bias identifiable, and sinks the value bias; mixture-of-experts layers add expert permutations. For these variants we determine the exact gauges, not the full fibre. For the residual stream, normalization placement is decisive: pre-norm models admit any orthogonal change of residual basis as an exact symmetry, including at $\varepsilon > 0$, so residual coordinates carry no functional meaning; models that normalize sublayer outputs (OLMo-2) admit only signed permutations (Figure 1b). A global rescaling of the residual stream is a symmetry only at $\varepsilon = 0$.

- **Verification on real weights at scale** (Section 5.1, Figure 2). We apply every predicted symmetry and a matched negative control directly to the trained weights of seven open model families from 124M to 32B parameters, including two mixture-of-experts models, in float64, on natural text, random tokens and Gaussian hidden states. Doing so required fixing a measurement pitfall: common implementations silently round float64 models to float32 inside normalizations, which makes different weights produce bit-identical logits.
- **What replicates across training runs** (Section 5.2, Figure 3). On independently trained seeds of two families (Pythia-160M and 410M, ten seeds each; GPT-2 small and medium, five each), head-level IOI and induction attributions are exactly equivariant within a checkpoint, carry no information across seeds when compared by head index (rank correlation at chance level), and replicate when heads are matched by symmetry-invariant signatures (attention patterns, or in GPT-2 circuits computed from the weights alone) or compared through label-invariant statistics.

We introduce no new interpretability method; the contribution is the symmetry analysis, its verification, and the resulting criterion (Section 4.4). Section 3 fixes notation and the three kinds of claims, Section 4 states the results, and Section 5 tests them.

2 Related Work

Parameter symmetries and identifiability. Hidden units of feedforward networks can be permuted, and rescaled (homogeneous activations) or sign-flipped (odd activations), without changing the function (Hecht-Nielsen, 1990); for generic tanh networks these are the only symmetries (Sussmann, 1992; Fefferman, 1994), while ReLU networks can have additional hidden ones (Grigsby et al., 2023). Symmetries matter for mode connectivity and model merging (Entezari et al., 2022; Ainsworth et al., 2023), including rotations inside attention (Zhang et al., 2025), and propagate to internal representations (Godfrey et al., 2022). For attention, Elhage et al. (2021) observed informally that only the QK and OV products matter; the key bias is redundant in standard attention (Namazifar et al., 2023). Concurrently with this work, Morrow (2026) derives the per-head gauges and their reduction by RoPE and grouped-query attention on a toy model, and Kim & Kim (2026) learn multi-head attention from queries, recovering the same canonical form as our Theorem 4.1; neither treats the residual stream, normalization placement, QK-normalization, attention sinks or mixture-of-experts, nor verifies on pretrained models, and we show that the key bias is *not* redundant under RoPE or with sinks. Brunner et al. (2020) study a different question, the identifiability of attention *weights* given the value vectors.

Residual-stream bases. For pre-norm transformers, folding the normalization gains makes the network invariant to orthogonal changes of the residual basis (Ashkboos et al., 2024a), a fact exploited for compression and quantization (Ashkboos et al., 2024b; Liu et al., 2025). Trained models nevertheless show basis-aligned phenomena such as outlier dimensions and massive activations (Elhage et al., 2023; Sun et al., 2024).

Reliability of mechanistic claims. Explanations of a fixed behavior can be non-unique (Méloux et al., 2025), subspace patching can create illusions (Makelov et al., 2024), and patching results depend on methodology (Zhang & Nanda, 2024). Circuits and features vary across training runs: learned group-operation circuits are arbitrary among a family (Chughtai et al., 2023), sparse autoencoders trained with different seeds share a minority of features (Paulo & Belrose, 2025), and pretraining seeds differ in stability (van der Wal et al., 2025). Representations of many discriminative models are identifiable only up to linear maps (Roeder et al., 2021), which motivates studying concept directions (Park et al., 2024) through invariants.

3 PRELIMINARIES

Attention layers. A multi-head attention layer with H heads maps $X \in \mathbb{R}^{n\times d}$ (one row per token) to

$$\mathrm{MHA}_\theta(X) = \sum_{h=1}^{H} \mathrm{softmax}\big(XW_h^Q(XW_h^K)^\top/\sqrt{d_k} + C\big)\, XW_h^V W_h^O,$$

with $W_h^Q, W_h^K \in \mathbb{R}^{d\times d_k}$, $W_h^V \in \mathbb{R}^{d\times d_v}$, $W_h^O \in \mathbb{R}^{d_v\times d}$ the head's row-block of the output projection, softmax taken row-wise and C either 0 or the causal mask. We write $M_h := W_h^Q(W_h^K)^\top/\sqrt{d_k}$ and $N_h := W_h^V W_h^O$ for the head's query–key form and value–output map (Elhage et al., 2021), and

$$\mu_\theta := \sum_h \delta_{M_h} N_h, \qquad \text{i.e.} \quad \mu_\theta(A) = \sum_{h:\, M_h = A} N_h,$$

for the finitely supported measure that assigns to each query–key form the total value–output map of the heads that use it. Let $G := \big(GL(d_k)^H \times GL(d_v)^H\big) \rtimes S_H$ act by relabelling heads and by the per-head gauges $(W_h^Q, W_h^K) \mapsto (W_h^Q S_h, W_h^K S_h^{-\top})$ and $(W_h^V, W_h^O) \mapsto (W_h^V A_h, A_h^{-1} W_h^O)$.

Proposition 3.1 (Head permutation and per-head gauges)**.** $\mathrm{MHA}_\theta(X) = \sum_h \mathrm{softmax}(XM_hX^\top + C)\, XN_h$*, so the layer depends on* θ *only through* $\{(M_h, N_h)\}_{h\le H}$*, in particular only through* μ_θ*, and is invariant under* G*.*

Residual stream. A transformer adds the outputs of its sublayers to a residual stream $x \in \mathbb{R}^d$. We call *readers* the linear maps applied to normalized residual activations (W^Q, W^K, W^V, MLP input and router weights, the unembedding) and *writers* the maps whose output is added to it (W^O, the MLP output, the embeddings). Normalization is placed either before each sublayer (pre-norm: GPT-2, Pythia, Llama, Qwen, gpt-oss) or on each sublayer's output (OLMo-2). A change of residual basis $x \mapsto xT$ is realized by a parameter change if readers can absorb T^{-1} and writers emit T. We write $\mathrm{rms}(x) = (\|x\|_2^2/d + \varepsilon)^{1/2}$.

Identifiability. Two parameter settings are *equivalent* if they compute the same function; the *fibre* of θ is its equivalence class, and a symmetry group acts within fibres. A quantity $f(\theta)$ is *identifiable* if it is constant on fibres, and a *maximal invariant* of a group action is an invariant of which every other invariant is a function (Lehmann & Casella, 1998). When a group's orbits are exactly the fibres, its maximal invariant is precisely what the function determines. Otherwise a group only bounds it: what the group changes is not identifiable, what it preserves may not be either. We prove equality for an attention layer (Theorem 4.1); for a whole network our results are such bounds.

Three kinds of claims. Formally, a checkpoint claim $\mathcal{C}$ is well posed if $\mathcal{C}(g\cdot\theta) = g\cdot\mathcal{C}(\theta)$ for every symmetry g; a function claim must be a function of the maximal invariant; a training claim compares runs whose labels and bases are related by unknown symmetries (and also differ as functions), so it must be invariant or follow an explicit alignment, and replicate.

4 CONTRIBUTIONS

4.1 AN ATTENTION LAYER IS IDENTIFIED EXACTLY UP TO μ_θ

Proposition 3.1 is elementary; the question is whether G is *all* there is (Figure 1a).

Theorem 4.1 (Complete identifiability of an attention layer)**.** *For any two parameter settings* θ, θ'*, possibly with different numbers of heads and head dimensions, the following are equivalent: (i)* $\mathrm{MHA}_\theta(X) = \mathrm{MHA}_{\theta'}(X)$ *for all* $X \in \mathbb{R}^{3\times d}$*; (ii) the same holds for all* $X \in \mathbb{R}^{n\times d}$ *and all* n*; (iii)* $\mu_\theta = \mu_{\theta'}$*. This holds for causal and bidirectional attention.*

Proof idea (Appendix A.2). On the input $X = (v+w,\, v,\, v)$ the last output row equals $v^\top \sum_h N_h + \sum_h \sigma(v^\top M_h w - \ln 2)\, w^\top N_h$, with σ the logistic function. The duplicated token produces the shift $\ln 2$, which makes logistic ridge functions with distinct slopes linearly independent: their complex poles have real parts $\ln 2/\beta$ and cannot cancel. For generic w the slopes $M_h w$ separate the distinct

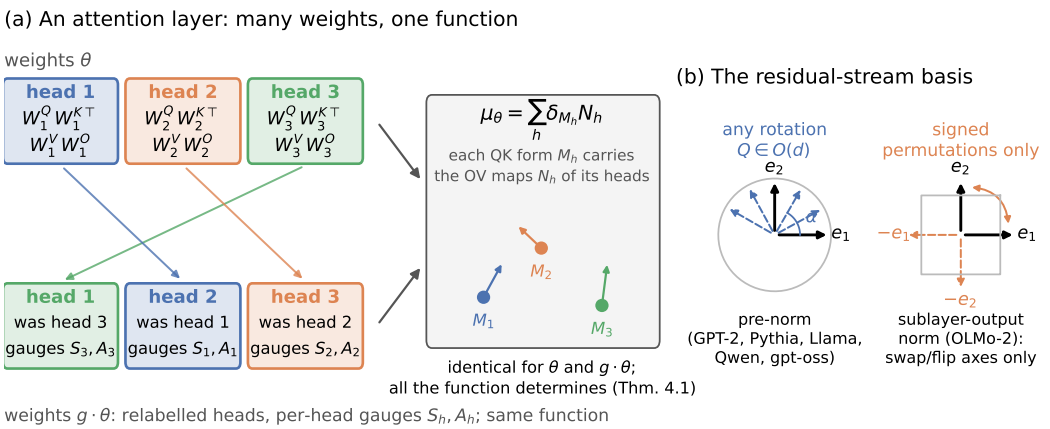

Figure 1: What the weights say and what the function determines. (a) Relabelling the heads of a layer and inserting $S_h S_h^{-1}$ or $A_h A_h^{-1}$ inside a head changes every weight matrix but not μ_θ, which is exactly what the layer's function determines (Theorem 4.1). (b) In a pre-norm model every rotation of the residual coordinates is an exact symmetry, so coordinates carry no functional meaning; normalizing sublayer outputs, as OLMo-2 does, leaves only signed permutations (Proposition 4.9).

forms, which isolates every atom of μ. Length two does not suffice: as $\sigma(-z) = 1 - \sigma(z)$, heads whose value–output maps sum to zero can jointly flip $(M_h, N_h) \mapsto (-M_h, -N_h)$. No genericity is assumed, and by analyticity any non-empty open set of length-three inputs suffices. The theorem concerns scores bilinear in the layer's inputs (positions, if any, enter through X, as in GPT-2); for RoPE and QK-normalization Section 4.2 gives the exact gauges, not completeness.

Corollary 4.2 (Completeness of G and the fibres)**.** *Let $d_k, d_v \leq d$ and let $\mathcal{U}$ be the set of parameters whose forms M_h are pairwise distinct, whose maps N_h are non-zero, and whose factors $W_h^Q, W_h^K, W_h^V, W_h^O$ have full rank. $\mathcal{U}$ is open and dense, and for $\theta \in \mathcal{U}$ the parameters of the same architecture computing the same function are exactly the orbit $G \cdot \theta$. Off $\mathcal{U}$ the fibre is $\{\theta' : \mu_{\theta'} = \mu_\theta\}$, which is strictly larger than the orbit: heads sharing a form can redistribute their value–output maps (e.g. by any invertible mixing of their joint value/output blocks), a head with $N_h = 0$ has an arbitrary form, and rank-deficient factors admit further factorizations.*

Hence μ_θ is the maximal invariant of the layer: a statement about the layer's function is well posed exactly when it is a function of μ_θ. Per-head statements such as "head h attends from names to names" are functions of μ_θ only up to relabelling (and, off $\mathcal{U}$, not even then). Kim & Kim (2026) recover the same canonical form as the output of a query-learning algorithm; Theorem 4.1 gives a short direct proof from length-three inputs and the fibre structure.

Are trained checkpoints generic? Genericity matters only for the completeness of G; every symmetry and measurement of Section 5 holds regardless. Of the nine checkpoints of Section 5 and the 30 of the cross-seed study, all but two lie inside $\mathcal{U}$ with a margin (Appendix B.4): no two query–key forms of a layer are closer than 14% in relative Frobenius distance, no head carries less than 15% of its layer's median value–output norm, and every factor has $\sigma_{\min}/\sigma_{\max} \geq 7 \times 10^{-5}$. Both OLMo-2 models are numerically non-generic: 12% (7B) and 9.5% (13B) of their query coordinates have weight columns below 10^{-6} of the largest and vanishing QK-normalization gains, which the function barely constrains: setting these gains to arbitrary values changes a layer's output by at most 10^{-8} of its update, against 0.008–0.6 for as many other coordinates.

4.2 Deployed architectures change the group

Each statement below is proved in Appendix A and tested in Section 5.1; unlike Theorem 4.1, they exhibit exact symmetries and exclude natural alternatives, without proving that no others exist.

Proposition 4.3 (Rotary embeddings)**.** *If RoPE (Su et al., 2024) rotates $r \leq d_k$ coordinates of each query and key in pairs with pairwise distinct frequencies in $(0, \pi)$, the query–key gauges that*

preserve all attention scores for generic weights are exactly $S_h \in (\mathbb{C}^\times)^{r/2} \times GL(d_k - r)$: a scaled rotation $\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)$ on each rotated pair and any invertible map on the unrotated coordinates.

Corollary 4.4 (Grouped-query attention)**.** *If query heads are partitioned into groups sharing one key and value head (Ainslie et al., 2023), the symmetries are permutations of query heads within a group, permutations of whole groups, and one query–key and one value–output gauge per group.*

Proposition 4.5 (QK-normalization moves the gauge into the gains)**.** *If queries and keys are RMS-normalized with learned gains before RoPE (Henry et al., 2020), the gauges of Proposition 4.3 no longer pass through the normalization: $n(yS) = n(y)S'$ for all y forces S to be a scaled orthogonal matrix. Instead, multiplying the query gains by any vector c that is constant on each rotated pair and dividing the key gains by c is an exact symmetry (one c per layer when gains are shared across heads, as in Qwen3; one per head when the normalization spans all heads, as in OLMo-2), as are sign flips of rotated pairs and, at $\varepsilon = 0$, positive rescalings of the query or key projection over the span of one normalization.*

Proposition 4.6 (Biases and attention sinks)**.** *(i) Shifting the key bias is an exact symmetry iff the shift lies in the unrotated coordinates and the softmax has no sink logit: RoPE turns the bias term into a position-dependent score, and a learned sink (Agarwal et al., 2025) is a logit that the shift does not move. (ii) Adding δ to a value bias and $-\sum_h \delta W_h^O$ (over the heads reading that value head) to the output bias is exact iff the attention weights sum to one (no sink).*

Part (i) refines the redundancy of the key bias (Namazifar et al., 2023; Morrow, 2026), which holds for GPT-2 but fails for RoPE models (Section 5.1). Finally, the experts of a mixture-of-experts layer can be permuted together with the router's rows, the neurons of any MLP can be permuted, and a SwiGLU neuron's linear branch can be rescaled by any non-zero factor, sign included, against its output weights; GELU and clamped gated units admit no such rescaling.

4.3 The residual stream: normalization placement decides the basis

Proposition 4.7 (Pre-norm: the function does not determine the residual basis)**.** *In a pre-norm transformer with RMSNorm, fold every normalization gain into the readers that follow it (and the final gain into an untied unembedding). Then for every orthogonal $Q \in O(d)$, replacing embeddings $E \mapsto EQ$, readers $W \mapsto Q^\top W$ and writers $W \mapsto WQ$ leaves the function unchanged for every $\varepsilon \geq 0$, and the residual stream of the new model is xQ at every layer. With LayerNorm, whose gain and bias are folded likewise, this holds as it stands for every Q that preserves the all-ones direction, since LayerNorm discards the mean; for every $Q \in O(d)$ it holds after centering all writers and the embeddings, which does not change the function, and replacing each LayerNorm by the RMS normalization it then computes, with the same ε (Ashkboos et al., 2024a).*

Proposition 4.7 separates two kinds of statements about residual coordinates. As statements about the computed function, they are not well posed in pre-norm models: "dimension 1415 carries a massive activation" can be rotated away without changing any output (Ashkboos et al., 2024b), and whatever the function determines is $O(d)$-invariant, i.e. a function of the Gram matrix of the residual vectors involved (Appendix A.6). As statements about the representative that training selected, they can be meaningful, since training is not rotation-invariant (Elhage et al., 2023; Sun et al., 2024), and the selection is strong: at the middle layer the median ratio $\max_i |x_i|/\mathrm{rms}(x)$ of residual states is 11.6–17.0 in GPT-2, SmolLM2 and Qwen2.5-7B, against 3.3–3.8 after a random rotation, which computes the same function (Appendix B.6); such claims concern training, to be tested across runs.

Proposition 4.8 (Scale: only at $\varepsilon = 0$, and only globally)**.** *At $\varepsilon = 0$, $\mathrm{RMSNorm}(\lambda x) = \mathrm{RMSNorm}(x)$ for $\lambda > 0$, and scaling the embeddings and all writers by λ is a symmetry; the direction $x/\|x\|$ is the maximal invariant of a norm input. Scaling a single layer's writer is not a symmetry, because the skip connection adds the unscaled stream. For $\varepsilon > 0$ the global rescaling changes the function by an amount governed by $\varepsilon/\mathrm{ms}(x)$ at the lowest-norm residual states.*

Deployed values of ε visibly break the rescaling (Section 5.1): the scale of the residual stream is weakly identifiable through ε alone, whereas rotations stay exact.

Proposition 4.9 (Sublayer-output normalization privileges the basis)**.** *Suppose every sublayer output is RMS-normalized with an elementwise gain before being added to the stream and the final layer applies RMSNorm (as in OLMo-2 (Team OLMo et al., 2025)), and that the gains of some sublayer norm have pairwise distinct absolute values. Then a change of residual basis $x \mapsto xT$ can be*

absorbed by linear changes of readers and writers and a change of gains iff T is a signed permutation matrix (or a positive multiple of one, when $\varepsilon = 0$). In the pre-norm setting of Proposition 4.7 the absorbable T are exactly the orthogonal matrices (and their positive multiples when $\varepsilon = 0$).

In OLMo-2, whose elementwise gains applied after normalization form a privileged basis, no change of basis other than reordering and sign flips preserves the function, so its residual coordinates cannot be rotated away (Figure 1b; nor can its median peak-to-RMS ratio of 19.4); conversely, OLMo-2 admits a per-layer rescaling of each writer, since its output is normalized before the addition. For the linear representation hypothesis (Park et al., 2024), if an observable depends on an activation only through its projection onto a subspace with orthonormal basis U, the maximal invariant of the basis is the projector $UU^\top$ (Appendix A.6): the subspace is identifiable, its individual axes are not unless an extra assumption such as sparsity selects them. In pre-norm models this indeterminacy is realized exactly by Proposition 4.7, for every subspace at once.

4.4 Consequences for interpretability claims

Checkpoint claims (Section 3) must be equivariant. "Head 9.6 moves names" passes: after relabelling, the same head, now numbered differently, still moves names, and activation patching reports exactly that (Section 5.2). Statistics of individual weight matrices fail: a function-preserving value–output gauge reverses the ranking of GPT-2's and Pythia's 144 heads by $\|W_h^O\|_F$ (Spearman $\rho = -1$), while the ranking by $\|N_h\|_F$ cannot move ($\rho = 1$; Appendix B.8). *Function claims* must be functions of μ_θ, of Gram matrices and projectors of residual vectors (pre-norm), or of the direction of a normalization input; a residual coordinate passes this test in OLMo-2, up to order and sign, but fails it in Qwen2.5. Inside a network the residual symmetries act on every head by conjugation, $(M_h, N_h) \mapsto (Q^\top M_h Q, Q^\top N_h Q)$ in a pre-norm model, so μ_θ is determined at best up to this action and head-level function claims must be invariant under it: attention patterns on the model's own activations, spectra of M_h and N_h, and circuits through its own embeddings are; raw entries of M_h are not. *Claims about what training produces* must replicate across runs, whose head labels are arbitrary relative to one another, like coordinates in an unknown basis; induction heads (Olsson et al., 2022) are the classic example (Section 5.2). The criterion is constructive (Appendix D): contract the indices on which gauges act (use M_h, N_h, not their factors), read residual directions through the model's own maps or activations, keep labels as pointers within a checkpoint and align or aggregate them across runs; and test any claim by recomputing it on weights transformed by random exact symmetries, which our released library does in one call.

5 Experiments

5.1 Measuring exact symmetries on trained weights

Protocol. For each model we load the pinned public checkpoint, apply an edit directly to the trained weight tensors, and compare the model's outputs with the unedited model on a 64-token English paragraph. Every predicted symmetry is paired with a negative control that differs from it minimally (a gauge applied per query head instead of per group, a rotation instead of a signed permutation, and so on). Models up to 7B parameters are run end to end in float64 and compared on logits. For the 13B–32B and mixture-of-experts models we load the model in bfloat16 on the GPU only to capture the real input of a decoder layer, copy that layer to float64, and compare its outputs; residual-stream symmetries are then checked as the per-layer equivariance $\mathrm{block}'(xT) = \mathrm{block}(x)T$, which composes across depth.

A measurement pitfall. Several Hugging Face implementations (Llama-family RMSNorm, some attention and router softmaxes) compute in float32 even in a float64 model: changing one weight of SmolLM2 by a relative 10^{-10}, or applying a dense value–output gauge, leaves the float64 logits *bit-identical* (Appendix B.1), so exact agreement is no evidence of a symmetry. We compute these operations in the model's precision and report, for every model, a *precision audit*: the logit change caused by a 10^{-10} relative change of one weight, which must be visible.

Results. We made 316 edits of the trained weights of nine models from seven families (Figure 2 shows the end-to-end runs and the middle layer of the larger models; Table 3 covers every tested

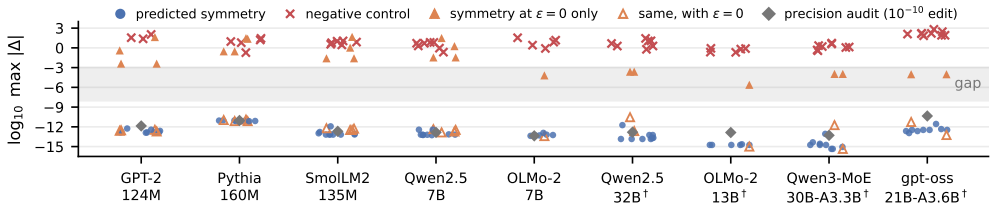

Figure 2: Every edit of the trained weights, one point each (Table 2). Predicted symmetries change the logits (†: the float64 output of a middle layer) only at float64 round-off, at the level of the precision audit; every negative control changes them by 0.1 to 2.5×10^3. Orange: global residual rescaling and, in OLMo-2, rescaling one layer's W^O, which are exact only at $\varepsilon = 0$ (hollow markers).

layer). Every predicted symmetry holds at float64 round-off, at the level of the precision audit and at least ten orders of magnitude below the smallest control of the same model; every control breaks. The architectural refinements are all visible (Table 2): the general query–key gauge breaks exactly in the RoPE models and the RoPE commutant repairs it, except under QK-normalization (OLMo-2, Qwen3), where only the exchange of gains survives; the key bias is free in GPT-2 and in Pythia's unrotated coordinates but not in rotated ones; the attention sinks of gpt-oss break the value-bias folding that is exact in GPT-2 and Pythia; SwiGLU neurons admit sign-changing rescalings that GELU and clamped units do not; the 128 experts of Qwen3-MoE and the 32 of gpt-oss can be permuted. For the residual stream, a global rotation is exact in every pre-norm model, from GPT-2 (after centering; Proposition 4.7) to the mixture-of-experts models, whereas OLMo-2 rejects it (a change of 34 logits) and accepts signed permutations; conversely, rescaling one layer's W^O is exact at $\varepsilon = 0$ only in OLMo-2. Global rescaling, exact at $\varepsilon = 0$, changes the logits by 0.004–0.3 ($\lambda = 10$) and up to 46 ($\lambda = 0.01$) at the deployed ε (Table 7). Repeating every run on random token sequences of 8 to 512 tokens and on Gaussian hidden states changes no conclusion (Appendix B.3).

5.2 Which head-level findings replicate across training runs?

Setup. We use two families of independently trained models: Pythia-160M and 410M (Biderman et al., 2023), ten seeds each on the same data (van der Wal et al., 2025), and GPT-2 small and medium, five seeds each on OpenWebText (Stanford CRFM, 2021). For every seed we compute two standard head-level attributions: IOI importance, the fraction of the clean–corrupted logit difference recovered by patching one head's output into the corrupted run (Wang et al., 2023; Zhang & Nanda, 2024), and the induction score (Olsson et al., 2022) (Appendix C). To match heads across runs we use signatures invariant to every weight symmetry of Section 4 except relabelling (Section 4.4): a head's attention probabilities on 24 passages of natural text, and, from the weights alone, its direct-path circuits, i.e. M_h and N_h evaluated on a fixed set of token embeddings and unembeddings (Appendix C). For every pair of seeds we compare the attributions (i) by head index, within each layer, against a null of random within-layer relabellings; (ii) after matching heads within each layer by maximum-weight matching of their signatures, which never looks at the attributions; and (iii) through label-invariant per-layer statistics.

Results. Within a checkpoint, the attributions are exactly equivariant: relabelling the heads of a seed by a function-preserving permutation relabels its IOI importance matrix up to float32 round-off ($\leq 3 \times 10^{-5}$), while comparing the two matrices by index gives $\rho = 0.08$–0.13. Across seeds, index-matched attributions agree at chance level in both families, both sizes and both scores (within-layer ρ between -0.02 and 0.02, below the 95% null quantile of 0.09–0.15): the index of a head carries no information about the head of another run. The same attributions replicate once heads are matched by attention signatures ($\rho = 0.41$–0.74 for induction, 0.11–0.31 for IOI; top-10 overlap rises from 0.04–0.22 to 0.14–0.51), and per-layer maxima agree ($\rho = 0.48$–0.77; Figure 3). In GPT-2, matching by weight circuits alone does almost as well ($\rho = 0.57$–0.74 and 0.17–0.35) and pairs a head with the same partner as the attention match for 46–67% of heads (chance: 6–8%); in Pythia these circuits are close to chance and weight spectra do better (induction 0.32–0.44), so which

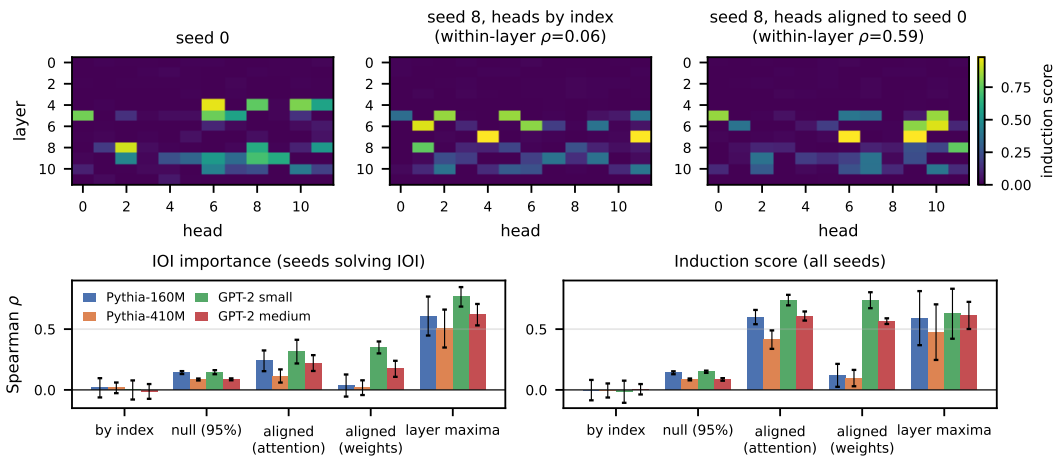

Figure 3: Head-level findings across training runs. Top: induction scores of two independently trained Pythia-160M seeds. Compared by index, seed 8 looks unrelated to seed 0; after its heads are aligned to seed 0's within each layer by their attention patterns on natural text, the induction heads line up (seed 8 is the median case among the nine seeds). Bottom: agreement over pairs of seeds (mean $\pm$ s.d.; 45 Pythia pairs, 36 for IOI; 10 GPT-2 pairs); "null": 95% quantile under random within-layer relabellings; weights: matched by direct-path circuits.

weight invariant best identifies a head across runs depends on the model. Induction heads replicate better than IOI heads, and larger models' heads less. Two Pythia seeds below our IOI accuracy threshold of 0.8 (0.45 and 0.79; the others 0.84–1.00) are excluded from the IOI comparisons, which changes no conclusion (Appendix C). A claim that "models of this family contain head (l, h) with role X" is therefore unfounded, while "models of this family contain, in layer l, a head whose attention pattern is P and whose role is X" is testable and, for induction heads, largely true.

6 Limitations and Conclusion

Limitations. Theorem 4.1 and Corollary 4.2 concern one attention layer with bilinear scores; with rotary embeddings or QK-normalization we give the exact gauges, not the full fibre, and for a whole network we exhibit exact symmetries without proving them complete. Restricting the inputs can only enlarge fibres: every symmetry we list holds on any input set, whereas identifying μ_θ needs an open set of inputs. A trained network supplies only its own normalized states; on them LayerNorm adds exact symmetries, which in a small trained model are the only local ones (Appendix B.5), but a global statement is open. Our statements concern the computed function; training prefers basis-aligned representatives (Appendix B.6), a property of training to be tested across runs, not of the function of a pre-norm model. Real checkpoints can lie near the non-generic set (OLMo-2). The cross-seed study covers two families, four sizes and two attributions, so its aligned correlations are lower bounds on what replicates; models above 7B parameters were verified layer by layer.

Conclusion. The function of a transformer determines much less than its parameters: exactly μ_θ for an attention layer and, for the residual stream, at most inner products in a pre-norm model, where every rotation preserves the function, whereas sublayer-output normalization admits only reordering and sign flips of the basis. These answers depend on the architecture, as we verified on seven model families up to 32B parameters. For interpretability the consequence is a checkable criterion rather than a prohibition: a claim about a checkpoint must be equivariant, a claim about a function invariant, and a claim about what training produces must survive independent runs, which head indices do not and label-invariant or aligned statistics do.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, et al. Smollm2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations (ICLR)*, 2024a.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024b.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *International Conference on Learning Representations (ICLR)*, 2020.

Lewis Carroll. *Alice's Adventures in Wonderland*. Macmillan, 1865. Public domain; text from Project Gutenberg.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Ronen Eldan and Yuanzhi Li. TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL `https://transformer-circuits.pub/2021/framework/index.html`.

Nelson Elhage, Robert Lasenby, and Christopher Olah. Privileged bases in the transformer residual stream. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/privileged-basis/index.html`.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Charles Fefferman. Reconstructing a neural net from its output. *Revista Matemática Iberoamericana*, 10(3):507–555, 1994.

Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.

J. Elisenda Grigsby, Kathryn Lindsey, and David Rolnick. Hidden symmetries of ReLU networks. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Robert Hecht-Nielsen. On the algebraic structure of feedforward network weight spaces. In R. Eckmiller (ed.), *Advanced Neural Computers*, pp. 129–135. Elsevier North-Holland, 1990.

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

Sunyeop Kim and Insung Kim. Provably learning multi-head attention with queries. *arXiv preprint arXiv:2608.03294*, 2026.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, New York, 2nd edition, 1998.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *International Conference on Learning Representations (ICLR)*, 2025.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations (ICLR)*, 2025.

W. Ross Morrow. Multi-head self attention is a parameter identification mechanism. *arXiv preprint arXiv:2609.01231*, 2026.

Mahdi Namazifar, Devamanyu Hazarika, and Dilek Hakkani-Tür. Role of bias terms in dot-product attention. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL `https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html`.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 39643–39666, 2024.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Geoffrey Roeder, Luke Metz, and Diederik P. Kingma. On linear identifiability of learned representations. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Stanford CRFM. Mistral: A framework for transparent and accessible large-scale language model training. https://github.com/stanford-crfm/mistral, 2021.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling (COLM)*, 2024.

Héctor J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4):589–593, 1992.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, et al. 2 OLMo 2 furious. In *Conference on Language Modeling (COLM)*, 2025.

Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. PolyPythias: Stability and outliers across fifty language model pre-training runs. In *International Conference on Learning Representations (ICLR)*, 2025.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Binchi Zhang, Zaiyi Zheng, Zhengzhang Chen, and Jundong Li. Beyond the permutation symmetry of transformers: The role of rotation for model fusion. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024.

A PROOFS

Throughout, vectors are rows, $\sigma(z) = 1/(1+e^{-z})$, and for a parameter setting θ of an attention layer $M_h = W_h^Q (W_h^K)^\top / \sqrt{d_k}$, $N_h = W_h^V W_h^O$ and $\mu_\theta(A) = \sum_{h: M_h = A} N_h$.

A.1 PROPOSITION 3.1

Row i of head h's output is $\sum_j p_{ij}^h \, x_j W_h^V W_h^O$ with $p_{i\cdot}^h = \mathrm{softmax}_j(x_i W_h^Q (W_h^K)^\top x_j^\top / \sqrt{d_k} + C_{ij})$, which depends on the weights only through M_h and N_h. The gauges fix M_h and N_h ($W_h^Q S \, (W_h^K S^{-\top})^\top = W_h^Q (W_h^K)^\top$ and $W_h^V A \, A^{-1} W_h^O = W_h^V W_h^O$), and relabelling heads permutes the terms of a sum. □

A.2 THEOREM 4.1

Lemma A.1. *Let $\beta_1, \dots, \beta_m$ be pairwise distinct reals and $c_1, \dots, c_m \in \mathbb{R}$. If $\sum_i c_i \, \sigma(\beta_i t - \ln 2) = 0$ for all $t \in \mathbb{R}$, then $c_1 = \cdots = c_m = 0$.*

Proof. For $\beta \neq 0$, $f_\beta(t) = \sigma(\beta t - \ln 2) = (1 + 2e^{-\beta t})^{-1}$ extends to a meromorphic function on $\mathbb{C}$ whose poles are the solutions of $e^{-\beta t} = -\frac{1}{2}$, namely $t_k = (\ln 2 + i\pi(2k+1))/\beta$, $k \in \mathbb{Z}$. They are simple, with residue $1/\beta$ (the derivative of $1 + 2e^{-\beta t}$ equals β there). Since $\mathrm{Re}\, t_k = \ln 2/\beta$, distinct non-zero slopes have disjoint pole sets. For $\beta = 0$, $f_0 \equiv 1/3$. The function $g = \sum_i c_i f_{\beta_i}$ is meromorphic and vanishes on $\mathbb{R}$; as $\mathbb{C}$ minus a discrete set is connected, $g \equiv 0$ by the identity theorem. At a pole of f_{β_j} ($\beta_j \neq 0$) all other terms are holomorphic, so the residue of g there is c_j/β_j, which must vanish. Once all terms with $\beta_i \neq 0$ are removed, the remaining constant $c_i/3$ must vanish too. □

Proof of Theorem 4.1. (iii)⇒(ii): grouping heads with equal forms, $\mathrm{MHA}_\theta(X) = \sum_A \mathrm{softmax}(XAX^\top + C)\, X \mu_\theta(A)$, a sum over the finite support of μ_θ; the right side depends only on μ_θ. (ii)⇒(i) is immediate. (i)⇒(iii): fix $v, w \in \mathbb{R}^d$ and X with rows $v + w,\ v,\ v$. The last row attends to all three rows under both masks; for a form A its scores are $vA(v+w)^\top, vAv^\top, vAv^\top$, so the weight on the first row is $p = \sigma(vAw^\top - \ln 2)$ and the output of the corresponding head is $p(v+w)N + (1-p)vN = vN + p\, wN$. Summing over heads, the last output row is

$$F_\theta(v, w) = v \sum_A \mu_\theta(A) + \sum_A \sigma(vAw^\top - \ln 2)\, w \, \mu_\theta(A).$$

Assume $F_\theta = F_{\theta'}$. At $w = 0$ this gives $\sum_A \mu_\theta(A) = \sum_A \mu_{\theta'}(A)$. Let $\mathcal{A}$ be the union of both supports and $D = \mu_\theta - \mu_{\theta'}$; then $\sum_{A \in \mathcal{A}} \sigma(vAw^\top - \ln 2)\, wD(A) = 0$ for all v, w. Choose w with $(A - B)w^\top \neq 0$ for all distinct $A, B \in \mathcal{A} \cup \{0\}$ (the excluded set is a finite union of proper subspaces), so that the vectors $Aw^\top$ are pairwise distinct and non-zero for $A \neq 0$; then choose z outside finitely many hyperplanes so that the numbers $\beta_A = zAw^\top$ are pairwise distinct. Setting $v = tz$ gives $\sum_A \sigma(\beta_A t - \ln 2)\, wD(A) = 0$ for all t, and Lemma A.1, applied coordinatewise, yields $wD(A) = 0$ for all A. This holds for all w in a dense set, hence $D = 0$. □

Restricted inputs. Both sides of (i) are real-analytic functions of $X \in \mathbb{R}^{3 \times d}$ (compositions of polynomials, exponentials and division by a positive sum). If they agree on a non-empty open set of inputs, their difference vanishes on it and hence, by the identity theorem on the connected space $\mathbb{R}^{3 \times d}$, everywhere, so (iii) follows. Agreement on a set without interior, such as the hidden states a trained network produces, does not suffice in general; Appendix B.5 measures what it determines locally in a trained model.

Why length three. With inputs $(v + w, v)$ of length two a head contributes $vN + \sigma(vAw^\top)\, wN$ to the last row. Since $\sigma(-z) = 1 - \sigma(z)$, replacing (A, N) by $(-A, -N)$ changes this contribution by $-(2v + w)N$, so flipping jointly every head of a set whose value–output maps sum to zero leaves all length-two outputs unchanged, although it changes μ_θ. The duplicated token of the length-three input shifts the argument by $\ln 2$, which breaks this symmetry (Lemma A.1).

A.3 COROLLARY 4.2

Lemma A.2 (Factorizations). *If $AB^\top = A'B'^\top$ with $A, B, A', B' \in \mathbb{R}^{d\times k}$ and A, B of rank k, then $A' = AS$ and $B' = BS^{-\top}$ for a unique $S \in GL(k)$.*

Proof. $B^\top$ has full row rank, so $\mathrm{col}(AB^\top) = \mathrm{col}(A)$ has dimension k; it is contained in $\mathrm{col}(A')$, which has dimension at most k, hence equal. Thus $A' = AS$ with S invertible, and $ASB'^\top = AB^\top$ with A of full column rank gives $SB'^\top = B^\top$. □

Proof of Corollary 4.2. The conditions defining $\mathcal{U}$ fail on a finite union of proper algebraic subsets, so $\mathcal{U}$ is open and dense. Let $\theta \in \mathcal{U}$ and let θ' (same architecture) compute the same function. By Theorem 4.1 $\mu_{\theta'} = \mu_\theta$, which has exactly H atoms $M_1, \dots, M_H$ with non-zero weights. Each atom must be the form of at least one head of θ', which has H heads, so there is a permutation π with $M'_{\pi(h)} = M_h$ and $N'_{\pi(h)} = N_h$. Since M_h has rank d_k (both factors have rank d_k), Lemma A.2 applies to (W_h^Q, W_h^K) and $(W'^Q_{\pi(h)}, W'^K_{\pi(h)})$, and likewise to $(W_h^V, (W_h^O)^\top)$, giving the per-head gauges.

Off $\mathcal{U}$ the fibre is $\{\theta' : \mu_{\theta'} = \mu_\theta\}$ by Theorem 4.1. Every element of the orbit $G \cdot \theta$ has the same multiset of pairs $\{(M_h, N_h)\}$ as θ, so it suffices to exhibit, in each non-generic case, a θ' with $\mu_{\theta'} = \mu_\theta$ and a different multiset. (a) Two heads share a form, $M_1 = M_2$, and $(N_1, N_2) \neq (0, 0)$: replacing their joint blocks $[W_1^V\ W_2^V]$ and $[W_1^O; W_2^O]$ by $[W_1^V\ W_2^V]B$ and $B^{-1}[W_1^O; W_2^O]$, $B \in GL(2d_v)$, preserves $N_1 + N_2$ and hence μ_θ; the new maps N_1', N_2' depend polynomially on B and B^{-1} and equal N_1, N_2 only on a proper algebraic subset, so a generic B changes the multiset. (b) A head has $N_h = 0$: it does not contribute to μ_θ, so replacing its form by any matrix that is not a form of θ changes the multiset. (c) A factor is rank-deficient, say $W_h^Q u = 0$ with $u \neq 0$ (the other factors are analogous), and $d_k < d$: then $W_h^K + vu^\top$ gives the same M_h for every $v \in \mathbb{R}^d$. The gauges act on W_h^K by right multiplication, which preserves its column space, and permutations only relabel heads, so the multiset of column spaces $\{\mathrm{col}(W_h^K)\}_h$ is invariant along the orbit. For $v \notin \mathrm{col}(W_h^K)$ the column space of $W_h^K + vu^\top$ contains $W_h^K u + \|u\|^2 v \notin \mathrm{col}(W_h^K)$, so for generic v this multiset changes and the new parameters leave the orbit. Numerical confirmations of (a) and (b) are in Appendix B.9; case (c) occurs in the OLMo-2 checkpoints (Appendix B.4). □

A.4 ARCHITECTURAL REFINEMENTS

Proof of Proposition 4.3. RoPE multiplies a query at position i by $R(i)$ and a key at j by $R(j)$, block-diagonal with rotations by $i\omega_\ell$ on the rotated pairs and identity elsewhere, so the score is $q\,R(i-j)\,k^\top$ with $R(\Delta) = R(i)R(j)^\top$. Under the gauge $(q, k) \mapsto (qS, kS^{-\top})$ it becomes $qSR(\Delta)S^{-1}k^\top$, which equals the original for all queries and keys in open sets (generic weights of rank d_k) iff S commutes with $R(\Delta)$ for every occurring offset, in particular $\Delta = 1$. The eigenvalues of $R(1)$ are $e^{\pm i\omega_\ell}$, pairwise distinct and different from 1 because $\omega_\ell \in (0, \pi)$, and 1 on the unrotated coordinates. A matrix commuting with $R(1)$ preserves its invariant subspaces: each rotated plane, on which it must commute with a rotation by $\omega_\ell \notin \pi\mathbb{Z}$ and hence be of the form $aI + bJ$, and the unrotated subspace, on which it is arbitrary. Conversely such matrices commute with every $R(\Delta)$. The standard frequencies $\omega_\ell = b^{-2\ell/r}$, also after linear or YaRN rescaling, lie in $(0, 1]$ and are pairwise distinct. □

Proof of Corollary 4.4. A key (value) head shared by several query heads can absorb only one gauge, so the gauges of the query heads of a group must agree (for W_g^K of full rank, $W_g^K S^{-\top} = W_g^K S'^{-\top}$ forces $S = S'$); permutations must keep each query head attached to its key/value head. The listed transformations satisfy both. □

Proof of Proposition 4.5. Write $q = \gamma_q \odot n(xW^Q)$ and $k = \gamma_k \odot n(xW^K)$ with n the gain-free RMS normalization over the normalized block and γ_q, γ_k the gains; the score is $n(xW^Q)\,\mathrm{diag}(\gamma_q)R(\Delta)\mathrm{diag}(\gamma_k)\,n(xW^K)^\top$. *Gauges.* If $n(yS) = n(y)S'$ for all y in an open set, then at $\varepsilon = 0$ $yS\,\mathrm{rms}(y) = yS'\,\mathrm{rms}(yS)$, so yS and yS' are parallel for all such y; as in the proof of Proposition 4.9, this forces $S' = \kappa S$ and $\mathrm{rms}(yS) = \kappa^{-1}\mathrm{rms}(y)$, i.e. S is a scaled

orthogonal matrix. *Exact reparametrizations.* If c is constant on every rotated pair, $\mathrm{diag}(c)$ commutes with $R(\Delta)$, so $\gamma_q \mapsto \gamma_q \odot c$, $\gamma_k \mapsto \gamma_k \oslash c$ leaves every score unchanged; c must be shared by all heads that share a gain vector or a key head. For a diagonal sign matrix D that is constant on rotated pairs, $n(yD) = n(y)D$, D commutes with the gains and with $R(\Delta)$, and $D^2 = I$, so $(W^Q, W^K) \mapsto (W^Q D, W^K D)$ is exact. At $\varepsilon = 0$, $n(\lambda y) = n(y)$ for $\lambda > 0$ gives the rescalings. That the general gauges of Proposition 4.3 break the function is verified on all three QK-normalized models (Table 2). □

Proof of Proposition 4.6. (i) With biases the score between positions i and j is $(x_i W^Q + b_Q)R(i-j)(x_j W^K + b_K)^\top$; shifting b_K by δ adds $(x_i W^Q + b_Q)R(i-j)\delta^\top$, which is constant in j, hence absorbed by the softmax, iff $R(\Delta)\delta^\top = \delta^\top$ for all offsets, i.e. δ is supported on unrotated coordinates, and there is no sink logit (a sink is concatenated to the scores and is not shifted, so a common shift changes the sink probability). For generic weights these conditions are also necessary. (ii) The head output is $\sum_j p_{ij}(x_j W^V + b_V)W^O = \sum_j p_{ij} x_j W^V W^O + (\sum_j p_{ij})\, b_V W^O$. If $\sum_j p_{ij} = 1$ the last term is a constant that the output bias absorbs; with a sink $\sum_j p_{ij} < 1$ varies with the input. □

MLPs and experts. Permuting hidden units (with the corresponding rows of input and columns of output weights) or experts (with the router's rows and biases) permutes the terms of a sum or of an equivariant top-k selection. For SwiGLU, $\mathrm{silu}(xW_g) \odot (xW_u)W_d$ is linear in the n-th column of W_u, so scaling it by $c_n \neq 0$ and the n-th row of W_d by $1/c_n$ is exact; GELU and the clamped gate of gpt-oss are not homogeneous.

A.5 Residual stream

Proof of Proposition 4.7. After folding, every normalization is $n(x) = x/\mathrm{rms}(x)$, and $\mathrm{rms}(xQ) = \mathrm{rms}(x)$ for orthogonal Q and every ε, so $n(xQ) = n(x)Q$. If the residual stream entering a block is xQ, its readers compute $n(xQ)Q^\top W = n(x)W$, its writers output yQ, and the stream leaving it is $(x+y)Q$; by induction from the embeddings, and at the end $n(x_L Q)Q^\top W_U^\top = n(x_L)W_U^\top$. For LayerNorm, $\mathrm{LN}(x) = n(xC)\,\mathrm{diag}(\gamma) + \beta$ with $C = I - \frac{1}{d}\mathbf{1}^\top\mathbf{1}$ (for row vectors $\mathbf{1}$), and ε enters n exactly as in RMSNorm since $\|xC\|^2/d$ is the variance; after folding γ into the readers and β into their biases, the readers see $n(xC)$. If Q is orthogonal with $\mathbf{1}Q = \pm\mathbf{1}$, then Q commutes with $\mathbf{1}^\top\mathbf{1}$ and hence with C, so $n(xQC) = n(xC)Q$ and the argument above applies to the LayerNorm model as it stands. For general Q, centering every writer and the embeddings ($W \mapsto WC$) changes the residual stream only by multiples of $\mathbf{1}$, which every LayerNorm removes, so the function is unchanged; on the resulting centered stream LayerNorm computes n followed by its affine map, so it can be replaced by n with the affine map folded into the readers (biases included), after which the argument above applies. □

Proof of Proposition 4.9. Readers can absorb any invertible T^{-1} since they are linear. A sublayer writes $y = n(a)\,\mathrm{diag}(\gamma)$, where a is its pre-normalization output, and for generic weights a ranges over an open set; the new model must write $yT = n(aS)\,\mathrm{diag}(\gamma')$ for a linear change S of the writer and new gains γ'. At $\varepsilon = 0$ this reads $a\,P\,\mathrm{rms}(aS) = a\,P'\,\mathrm{rms}(a)$ with $P = \mathrm{diag}(\gamma)T$ and $P' = S\,\mathrm{diag}(\gamma')$, both invertible. Hence $b\,P^{-1}P'$ is parallel to b for every b in the open set $\{aP\}$, so every such b is an eigenvector and $P^{-1}P' = \kappa^{-1}I$: $\mathrm{diag}(\gamma)T = \kappa\,S\,\mathrm{diag}(\gamma')$ and $\mathrm{rms}(aS) = \kappa^{-1}\mathrm{rms}(a)$, i.e. $S = \kappa^{-1}O$ with O orthogonal. At the final RMSNorm the logits are $n(x)\,L$ with $L = \mathrm{diag}(\gamma_f)W_U^\top$ of full row rank, and must equal $n(xT)\,L'$: then $xTL' = \lambda(x)\,xL$ for x in an open set; comparing x_1, x_2 and $x_1 + x_2$, with $x_1 L, x_2 L$ linearly independent, shows that λ is constant, so $TL' = \lambda L$ and $\mathrm{rms}(xT) = \lambda\,\mathrm{rms}(x)$, i.e. $T = cO''$ with $c > 0$ and O'' orthogonal. Combining, $A := \mathrm{diag}(\gamma)O'' = c^{-1}O\,\mathrm{diag}(\gamma')$ satisfies $A^\top A = O''^\top \mathrm{diag}(\gamma^2)O'' = c^{-2}\mathrm{diag}(\gamma'^2)$: O'' diagonalizes $\mathrm{diag}(\gamma^2)$, whose eigenvalues are distinct, so O'' is a signed permutation. For $\varepsilon > 0$, $\mathrm{rms}_\varepsilon(xcO'') = \lambda\,\mathrm{rms}_\varepsilon(x)$ for all x forces, at $x = 0$, $\lambda = 1$ and then $c = 1$. Conversely, for a signed permutation P, $n(aP) = n(a)P$ at every ε and $(n(a)\mathrm{diag}(\gamma))P = n(aP)\mathrm{diag}(\gamma_P)$ with γ_P the permuted gains, and a factor $c > 0$ is absorbed by the gains and the final normalization at $\varepsilon = 0$. For pre-norm models the final-norm argument alone gives $T = cO''$ (and $c = 1$ for $\varepsilon > 0$), and every orthogonal O'' is absorbable (Proposition 4.7). □

Proof of Proposition 4.8. At $\varepsilon = 0$, $\mathrm{rms}(\lambda x) = |\lambda|\,\mathrm{rms}(x)$, so $n(\lambda x) = \mathrm{sign}(\lambda)\,n(x)$, and by the induction of Proposition 4.7 (with Q replaced by λI) the global rescaling is a symmetry for $\lambda > 0$. Rescaling one writer adds λy to an unscaled x, and $x + \lambda y$ is not proportional to $x + y$ in general. For $\varepsilon > 0$, $n_\varepsilon(\lambda x)/n_\varepsilon(x) = \lambda\big((\mathrm{ms}(x) + \varepsilon)/(\lambda^2 \mathrm{ms}(x) + \varepsilon)\big)^{1/2}$ with $\mathrm{ms}(x) = \|x\|^2/d$, which differs from 1 by $O(\varepsilon|1 - \lambda^{-2}|/\mathrm{ms}(x))$. If f is invariant under $x \mapsto \lambda x$ for all $\lambda > 0$, then $f(x) = f(x/\|x\|)$, so the direction is the maximal invariant. □

A.6 Maximal invariants

Gram matrices. Two families $(x_1, \ldots, x_m)$ and $(x'_1, \ldots, x'_m)$ in $\mathbb{R}^d$ are related by some $Q \in O(d)$ iff they have the same Gram matrix: if $XX^\top = X'X'^\top$ for the stacked matrices, the map $x_i \mapsto x'_i$ is a well-defined isometry between the row spaces, which extends to an orthogonal map of $\mathbb{R}^d$. Hence a function of the family is $O(d)$-invariant iff it is a function of the Gram matrix. *Projectors.* For orthonormal $U, U' \in \mathbb{R}^{d\times k}$, $UU^\top = U'U'^\top$ iff $U' = UR$ with $R \in O(k)$: if the projectors agree the ranges agree, and $R = U^\top U'$ satisfies $R^\top R = U'^\top UU^\top U' = U'^\top U' = I$ and $UR = UU^\top U' = U'$. Hence $f(UR) = f(U)$ for all R iff f factors through $UU^\top$.

B Experimental Details

All experiments use the public checkpoints listed in Table 1, each pinned to a Hugging Face commit, with `transformers` 5.17 and PyTorch 2.14, on a single machine with an NVIDIA GB10 GPU and 128 GB of unified memory. Every number in the paper is produced by the released scripts, which write the raw results to JSON.

Model	Params	Attention	Norm (placement)	MLP	Mode
GPT-2 (Radford et al., 2019)	124M	MHA, learned pos., biases	LayerNorm (pre)	GELU	e2e
Pythia-160M (Biderman et al., 2023)	160M	MHA, RoPE on 25%, biases	LayerNorm (pre, parallel)	GELU	e2e
SmolLM2-135M (Allal et al., 2025)	135M	GQA 9/3, RoPE	RMSNorm (pre)	SwiGLU	e2e
Qwen2.5-7B (Yang et al., 2024)	7.6B	GQA 28/4, RoPE, QKV bias	RMSNorm (pre)	SwiGLU	e2e
OLMo-2-7B (Team OLMo et al., 2025)	7.3B	MHA, RoPE, QK-norm (all heads)	RMSNorm (sublayer outputs)	SwiGLU	e2e
Qwen2.5-32B (Yang et al., 2024)	32.8B	GQA 40/8, RoPE, QKV bias	RMSNorm (pre)	SwiGLU	layer
OLMo-2-13B (Team OLMo et al., 2025)	13.7B	MHA, RoPE, QK-norm (all heads)	RMSNorm (sublayer outputs)	SwiGLU	layer
Qwen3-30B-A3B (Yang et al., 2025)	30.5B	GQA 32/4, RoPE, QK-norm (per head)	RMSNorm (pre)	MoE 128/8	layer
gpt-oss-20b (Agarwal et al., 2025)	20.9B	GQA 64/8, RoPE, biases, sinks	RMSNorm (pre)	MoE 32/4	layer

Table 1: Models used in Table 2. e2e: whole model in float64, logits compared; layer: one middle decoder layer in float64 on its real inputs.

	GPT-2 124M	Pythia 160M	SmolLM2 135M	Qwen2.5 7B	OLMo-2 7B	Qwen2.5 32B†	OLMo-2 13B†	Qwen3-MoE 30B-A3.3B†	gpt-oss 21B-A3.6B†
Precision audit (Sec. 5.1)	-12	-11	-13	-13	-13	-13	-13	-13	-10
Heads permuted (in KV group)	-13	-11	-13	-13	-13	-14	-15	-15	-12
KV groups permuted	–	–	-13	-13	–	-13	–	-15	-12
ctrl: heads permuted, W^O kept	**35**	**6.2**	**7.5**	**5**	**15**	**4.7**	**0.84**	**1.3**	**159**
ctrl: query to other group	–	–	**10**	**6.7**	–	**12**	–	**1.6**	**163**
OV gauge $A_g \in GL(d_v)$	-13	-11	-13	-13	-13	-13	-15	-14	-12
QK gauge $S_g \in GL(d_k)$	-13	**9.2**	**6**	**5.1**	**7.5**	**11**	**0.64**	**5.3**	**289**
QK gauge, RoPE commutant	–	-11	-13	-13	**2.6**	-14	**0.21**	**2.2**	-12
ctrl: per-query-head gauge	–	–	**2.8**	**2**	–	**1.9**	–	**1.1**	**82**
QK-norm gain exchange	–	–	–	–	-13	–	-15	-15	–
Key bias shift, unrotated coords	-13	-11	–	–	–	–	–	–	–
Key bias shift, rotated coords	–	**0.19**	–	**0.22**	–	**0.6**	–	–	**69**
Value bias → output bias	-13	-11	–	–	–	–	–	–	**82**
MLP neurons permuted	-13	-11	-13	-13	-13	-14	-15	-15	-13
MLP neurons rescaled ($c_n<0$ ok)	**109**	**28**	-13	-13	-13	-14	-15	-15	**141**
Experts permuted (with router)	–	–	–	–	–	–	–	-15	-13
All of the above, every layer	-13	-11	-13	-12	-13	–	–	–	–
One layer's $W^O \times 3.7$	**24**	**16**	**11**	**6.8**	$7 \times 10^{-5}\,(-13)$	**27**	$2 \times 10^{-6}\,(-15)$	**3.9**	**651**
Residual rescale $\lambda = 10$	$4 \times 10^{-3}\,(-13)$	$0.3\,(-11)$	$0.03\,(-12)$	$0.04\,(-12)$	–	$2 \times 10^{-4}\,(-11)$	–	$10^{-4}\,(-12)$	$10 \times 10^{-5}\,(-11)$
Residual rotation $Q \in O(d)$	-12	-11	-12	-12	**34**	-12	**0.79**	-13	-12
Residual signed permutation	–	–	–	–	-13	–	-15	–	–

Table 2: Full symmetry atlas (the points of Figure 2). Predicted symmetries: exponent of the maximal absolute change of the logits (†: of the float64 output of the middle tested layer), e.g. -13 means $\sim 10^{-13}$. Predicted non-symmetries (controls): the change itself, in bold. Symmetries at $\varepsilon = 0$ only: change at the model's ε, and in parentheses at $\varepsilon = 0$. "–": not applicable. Every entry agrees with Section 4.

B.1 PRECISION

With the default Hugging Face implementation (float32 RMSNorm, SDPA attention), a float64 SmolLM2-135M returns bit-identical logits after a relative change of 10^{-10} to one weight of a middle layer, and after a dense value–output gauge with condition number 5 (maximal logit change 0.0 in both cases). With native-precision normalization and attention the same edits give 2.1×10^{-13} and 7.1×10^{-14}: the first is a real change, the second float64 round-off. A reported difference of exactly zero is therefore not evidence of an exact symmetry unless the measurement's sensitivity is shown. All our models are audited in this way (first row of Table 2). Transformers 5 also routes mixture-of-experts layers to grouped-matrix-multiply kernels without float64 support; we use the reference loop.

B.2 EDITS AND CONTROLS

Gauges are random with condition number exactly 5 (a fixed singular spectrum between two random orthogonal matrices); RoPE-commutant gauges draw, for each rotated pair, $a \in \pm[0.5, 2]$ and $b \in [-1, 1]$, and an invertible block on unrotated coordinates; permutations are drawn without the identity. For end-to-end runs, block-local edits are applied to the middle layer and undone exactly from a saved copy of its tensors, and the composed test applies permutations, both gauges (the gain exchange for QK-normalized models) and MLP permutations to every layer at once. Residual rotations use a Haar-random $Q \in O(d)$, after folding all gains (and, for LayerNorm models, centering all writers and replacing LayerNorm by a gain-free RMS normalization, which is exact on a centered stream). In layer mode the block's own gains are folded and the block is fed xQ; the reference is $\mathrm{block}(x)Q$.

Layer mode. For the larger models the full float64 model does not fit in memory. We run the model in bfloat16 once to capture, with a forward pre-hook, the exact arguments of the tested decoder layer (hidden states, attention mask, rotary cosines and sines); the layer is then copied to float64 and every block-local edit is evaluated on it. Because the residual stream carries massive activations, the scale used to classify a change is the size of the block's own update, $\max |\mathrm{block}(x) - x|$. To validate the procedure we ran it on Qwen2.5-7B, which we also verified end to end: on its first, middle and last layers all 48 verdicts agree with the end-to-end run. Table 3 lists every tested layer.

Model	layer	#exact	worst exact	#controls	smallest control	agree
OLMo-2-1124-13B	1	6	10^{-15}	5	0.17	yes
OLMo-2-1124-13B	20	6	2×10^{-15}	5	0.21	yes
OLMo-2-1124-13B	38	6	2×10^{-14}	5	1.9	yes
Qwen2.5-32B	1	9	8×10^{-15}	7	0.27	yes
Qwen2.5-32B	32	9	2×10^{-12}	7	0.6	yes
Qwen2.5-32B	62	9	5×10^{-12}	7	1.5	yes
Qwen2.5-7B	1	9	5×10^{-14}	7	0.87	yes
Qwen2.5-7B	14	9	7×10^{-13}	7	0.44	yes
Qwen2.5-7B	27	9	2×10^{-12}	7	1	yes
Qwen3-30B-A3B	1	9	7×10^{-14}	8	0.1	yes
Qwen3-30B-A3B	24	9	9×10^{-14}	8	0.37	yes
Qwen3-30B-A3B	46	9	2×10^{-13}	8	3	yes
gpt-oss-20b	1	9	2×10^{-13}	9	9.2	yes
gpt-oss-20b	11	9	3×10^{-12}	9	56	yes
gpt-oss-20b	12	9	3×10^{-12}	9	69	yes
gpt-oss-20b	22	9	7×10^{-12}	9	2.8×10^{2}	yes

Table 3: Layer-mode runs: for every tested layer, the number of predicted symmetries and the largest change among them, the number of controls and the smallest change among them, and whether every verdict agrees with the theory.

B.3 ROBUSTNESS TO THE INPUT

The atlas compares outputs on one English paragraph. We repeated every run with other inputs: uniformly random token ids (8, 128 and 512 tokens, end to end, for the three small models; 128 tokens for the 7B models and, in layer mode, for the larger ones), and Gaussian hidden states with the per-token RMS of the real ones, fed to the middle layer of every RMSNorm model, which tests

the algebra away from any data distribution. Table 4 reports, for every model and input, the margin between the smallest control and the largest deviation of a predicted symmetry. Over these 22 runs and 392 edits the margin never falls below 9.6 orders of magnitude (the OLMo-2 runs also contain the near-dead gain edit of Appendix B.4, which is neither a symmetry nor a control). One control falls below our classification threshold (10^{-4} of the logit scale): shifting Pythia's key bias in rotated coordinates changes the logits by 0.02 on 8 random tokens, against 0.19 on the paragraph and 0.37 on 128 tokens, because RoPE turns the shift into a score that grows with the distance between positions (Proposition 4.6); it is still 4×10^9 times the largest deviation of a predicted symmetry in that run.

Model	paragraph	random, 8	random, 128	random, 512	Gaussian states
GPT-2	13.6	13.3	13.3	13.3	–
Pythia-160M	10.3	9.6	10.0	10.2	–
SmolLM2-135M	12.4	11.9	12.2	12.4	13.1
Qwen2.5-7B	11.8	–	11.9	–	11.8
OLMo-2-7B	12.8	–	12.9	–	13.6
Qwen2.5-32B	11.5	–	11.0	–	11.4
OLMo-2-13B	13.9	–	12.5	–	12.4
Qwen3-30B-A3B	12.2	–	12.5	–	12.9
gpt-oss-20b	13.3	–	13.1	–	13.7

Table 4: Margin, in orders of magnitude, between the smallest control and the largest deviation of a predicted symmetry, for every input ("paragraph": the runs of Table 2; "random, n": n random tokens; Gaussian states: middle layer, RMSNorm models). "–": not run.

B.4 How generic are the checkpoints?

For every attention layer we read the per-head factors from the checkpoint and compute, with $M_h = W_h^Q (W_h^K)^\top / \sqrt{d_k}$ and $N_h = W_h^V W_h^O$, the relative separation $\|M_h - M_{h'}\|_F / (\|M_h\|_F + \|M_{h'}\|_F) \in [0, 1]$ of every pair of heads (from $d_k \times d_k$ Gram matrices, without forming the $d \times d$ forms), the smallest $\|N_h\|_F$ relative to the layer's median, and the smallest $\sigma_{\min}/\sigma_{\max}$ of $W_h^Q, W_h^K, W_h^V, W_h^O$ (Table 5). Pairs of heads in one key/value group share W^K but not M_h. For the QK-normalized models the normalization sits between projections and scores, so the forms M_h are a proxy. No column of any W_h^Q is exactly zero, but in OLMo-2-7B and 13B, 12.0% and 9.5% of all query coordinates have a column norm below 10^{-6} of the largest in their layer (median 10^{-16} and 2×10^{-10} of it), and their query-normalization gains are vanishing as well (median 1.4×10^{-8} and 2.4×10^{-13}, against 0.85 and 0.72 on the other coordinates); no other model has such coordinates, including Qwen3, whose QK-normalization acts per head. These parameters are almost unconstrained: setting the gains of the near-dead coordinates of a middle layer to values drawn from $[-3, 3]$ changes the layer's output by 4×10^{-10} to 1.1×10^{-8} of its update on random tokens and Gaussian states (and the logits of OLMo-2-7B by 1.9×10^{-7}), against 0.008 to 0.58 (and 1.1) when as many live coordinates are changed in the same way (Appendix B.3). The 30 checkpoints of the cross-seed study (Appendix C) are all generic, with minimal QK separation 0.14, OV mass 0.15 and $\sigma_{\min}/\sigma_{\max} = 7.6 \times 10^{-5}$, and none has near-dead coordinates.

Model	pairs	min QK separation	1% quantile	min OV mass	min $\sigma_{\min}/\sigma_{\max}$
GPT-2	792	0.51	0.53	0.49	2.5×10^{-2}
Pythia-160M	792	0.40	0.47	0.38	2.7×10^{-3}
SmolLM2-135M	1080	0.23	0.39	0.26	1.9×10^{-3}
Qwen2.5-7B	10584	0.35	0.58	0.49	6.8×10^{-3}
OLMo-2-7B*	15872	0.50	0.63	0.26	$< 10^{-10}$
Qwen2.5-32B	49920	0.32	0.61	0.45	3.1×10^{-4}
OLMo-2-13B*	31200	0.06	0.66	0.17	$< 10^{-10}$
Qwen3-30B-A3B*	23808	0.36	0.52	0.52	6.6×10^{-4}
gpt-oss-20b	48384	0.26	0.50	0.29	4.2×10^{-4}

Table 5: Distance of every layer from the non-generic set of Corollary 4.2, minimum over layers (and, for the separation, the 1% quantile over all pairs of heads). *: QK-normalized.

B.5 IDENTIFIABILITY ON THE STATES A MODEL REACHES

Theorem 4.1 assumes agreement on all inputs, whereas a trained network feeds each layer only the normalized hidden states it produces. We measured what these states determine, to first order, in a trained model whose attention is exactly the layer of Section 3: TinyStories-1M (Eldan & Li, 2023) (GPT-Neo, $d = 64$, 8 layers of 16 heads of dimension 4, no rotary embeddings). For every layer we compute in float64 the Jacobian J of the map from the $4d^2 = 16384$ weights (W^Q, W^K, W^V, W^O) to the layer's outputs on 512 input rows (4 sequences of 128 tokens). The gauge orbit contributes $H(d_k^2 + d_v^2) = 512$ null directions. On LayerNorm outputs there are more: every such state satisfies $a^\top x = c$ with $a = 1/\gamma$ and $c = \sum_i \beta_i/\gamma_i$, so replacing W_h^K by $W_h^K + ag^\top$ ($g \in \mathbb{R}^{d_k}$) adds $c\, x_i W_h^Q g$ to every score of query i, a constant that the softmax ignores: a key bias in disguise, and $Hd_k = 64$ further null directions. We test whether the null space of J is exactly spanned by the predicted directions: they must change the outputs only at round-off, and J must have full column rank on their orthogonal complement, which we check by a QR decomposition of J stacked with the predicted directions scaled by $\sigma_{\max}(J)$, whose smallest singular value (found by inverse iteration) equals the smallest singular value of J on that complement. On the model's own states, and on those of an unrelated book, the null space has dimension exactly 576 in every layer: the predicted directions change the outputs by at most 3×10^{-16} in relative terms, and on their complement the smallest singular value is 9×10^{-10} to 1.3×10^{-7} of the largest. On Gaussian rows, where the LayerNorm directions are not null (relative change 0.13–0.44), the null space is exactly the 512-dimensional tangent space of the gauge orbit, and 8 states, fewer equations than weights, leave 15872 null directions (Table 6). Locally, the states a trained model reaches thus determine its attention weights up to the gauges and the LayerNorm constraint and nothing else, but only weakly: the smallest other singular values are 10^{-9} to 10^{-7} of the largest on real states, against 10^{-4} on Gaussian inputs. (Our implementation reproduces the model's attention up to the float32 rounding of its scores, 1.4×10^{-6} in relative terms, and the real states satisfy $a^\top x = c$ to 2×10^{-15}.)

Layer	0	1	2	3	4	5	6	7
sampled stories	576	576	576	576	576	576	576	576
book passages	576	576	576	576	576	576	576	576
Gaussian rows	512	512	512	512	512	512	512	512
8 tokens only	15872	15872	15872	15872	15872	15872	15872	15872
smallest other $\sigma/\sigma_{\max}$, stories	$7{\cdot}10^{-8}$	$9{\cdot}10^{-10}$	$2{\cdot}10^{-9}$	$1{\cdot}10^{-8}$	$7{\cdot}10^{-8}$	$5{\cdot}10^{-9}$	$2{\cdot}10^{-8}$	$4{\cdot}10^{-8}$

Table 6: Dimension of the null space of the Jacobian of each attention layer of TinyStories-1M with respect to its weights (singular values below 10^{-13} of the largest), on 512 rows of: its own hidden states on stories it sampled, its hidden states on passages of an unrelated book, Gaussian rows, and 8 of its own hidden states. The gauges account for 512 null directions and the LayerNorm constraint for 64 more on real states. Last row: smallest singular value outside the null space, relative to the largest, on the sampled stories.

B.6 BASIS ALIGNMENT SELECTED BY TRAINING

On the 24 passages of Appendix C, we collect the residual states at the middle layer (excluding the first position) and compute $\max_i |x_i|/\mathrm{rms}(x)$ in the checkpoint's basis and after a Haar-random rotation, for which the ratio concentrates near $\sqrt{2 \ln d}$ (3.6–4.1 for these widths). Medians (99% quantiles): GPT-2 11.6 (15.4) against 3.3 (4.1); SmolLM2 13.0 (17.0) against 3.3 (4.3); Qwen2.5-7B 17.0 (27.4) against 3.8 (4.9); OLMo-2-7B 19.4 (39.9) against 3.8 (4.7). In the three pre-norm models the rotated states belong to a model computing the same function (Proposition 4.7); in OLMo-2 no such model exists (Proposition 4.9).

B.7 THE ε OF DEPLOYED MODELS BREAKS RESCALING

The change is concentrated where the residual stream is small: in the first layers the root-mean-square of residual states is 0.03–0.1 (Pythia-160M, SmolLM2), so $\varepsilon = 10^{-5}$ is not negligible against $\mathrm{ms}(x)$, and downscaling makes it dominant. Rotations leave $\mathrm{ms}(x)$ unchanged and are unaffected.

Model	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 10$	$\lambda = 100$
GPT-2 124M	44 (4×10^{-13})	0.41 (3×10^{-13})	4.0×10^{-3} (2×10^{-13})	4.1×10^{-3} (3×10^{-13})
Pythia 160M	30 (10^{-11})	24 (7×10^{-12})	0.3 (8×10^{-12})	0.3 (10^{-11})
SmolLM2 135M	46 (3×10^{-13})	1.1 (3×10^{-13})	0.026 (6×10^{-13})	0.026 (5×10^{-13})
Qwen2.5 7B	31 (10^{-13})	1.8 (2×10^{-13})	0.035 (4×10^{-13})	0.036 (3×10^{-13})

Table 7: Maximal logit change under a global residual rescaling by λ, at the model's ε and at $\varepsilon = 0$.

B.8 Weight statistics are not functional

For GPT-2, Pythia-160M and SmolLM2 we computed, for every head, the Frobenius norms of W_h^Q, W_h^V, W_h^O, the top singular value of W_h^V, and the same statistics of M_h and N_h. A random function-preserving gauge (condition number 10) leaves the logits unchanged to $\leq 10^{-11}$ and the statistics of M_h, N_h exactly unchanged, but perturbs the raw-weight statistics (rank correlation over heads down to 0.91 for the top singular value of W_h^V). A gauge chosen adversarially, a scalar a_h per head, reverses the ranking of heads by $\|W_h^O\|_F$ exactly in GPT-2 and Pythia (Spearman $\rho = -1$; logit change 1.4×10^{-13} and 8.4×10^{-12}) and yields $\rho = -0.44$ in SmolLM2, whose GQA gauge is shared by three heads, while the ranking by $\|N_h\|_F$ stays at $\rho = 1$.

B.9 Numerical check of Theorem 4.1

On random layers ($d = 12$, $d_k = d_v = 4$, $H = 3$, 200 draws) the closed form of the last output row on $(v + w, v, v)$ used in the proof matches the implementation to 2.1×10^{-14}. Two heads sharing a form, remixed by a random $B \in GL(8)$, compute the same function to 4.6×10^{-14} over random inputs of lengths 1–8, although their value–output maps change by 8.4 and are not related by G; a head with $N_h = 0$ can take an arbitrary form (difference 0); a gauge-and-permutation move changes the function by 1.1×10^{-13}, whereas perturbing one form by 10^{-6} changes it by 9.6×10^{-5}.

B.10 Synthetic validation

An independent numpy implementation of the definitions (vectorized and token-by-token versions agree to 1.2×10^{-14}) confirms head-permutation invariance to 3.4×10^{-13} against a floor of 16.1 for the broken variant (2880 trials, $H \in \{2, 4, 8, 16\}$, float64), RMSNorm scale invariance over 310 orders of magnitude of λ with an ε-induced deviation that scales linearly with ε. In random multi-layer stacks (300 trials per regime, $\varepsilon = 0$, pre-norm stacks ending with a normalization as real models do), rescaling each layer's W^O by its own factor breaks the function under both norm placements (median change 4.16 post-norm, 3.14 pre-norm), whereas the global rescaling of the input and all writers (maximal change 4.5×10^{-13}) and the global rotation with folded gains (2.4×10^{-12}) are exact, as are head permutations under both placements ($\leq 1.3 \times 10^{-12}$).

C Cross-Seed Study

Models. Pythia-160M and Pythia-410M from the PolyPythias suite (van der Wal et al., 2025): the original run (seed 0) and nine additional seeds each, trained on the same data (12 layers $\times$ 12 heads and 24 layers $\times$ 16 heads; LayerNorm, rotary embeddings on a quarter of each head, parallel attention and MLP). GPT-2 small and medium from the Mistral project (Stanford CRFM, 2021), trained on OpenWebText from seeds 21, 49, 81, 343 and 777 (12 $\times$ 12 and 24 $\times$ 16 heads; learned absolute positions, sequential blocks, a different tokenizer and different data). Final checkpoints, float32 on the GPU.

IOI importance. Eight templates in the style of Wang et al. (2023) (e.g. "When A and B went to the store, B gave a drink to"), each filled with the same 16 ordered pairs of single-token first names, give 128 prompts; the answer is A. The corrupted prompt replaces the second occurrence of B by A, which flips the answer and keeps the length. The metric is the logit difference $\mathrm{LD} = \ell(\mathrm{A}) - \ell(\mathrm{B})$ at the last position. For every head we cache its clean output (the input of the output projection, restricted to the head's block, at all positions), patch it into the corrupted run, and define its importance as $(\overline{\mathrm{LD}}_{\mathrm{patched}} - \overline{\mathrm{LD}}_{\mathrm{corrupt}})/(\overline{\mathrm{LD}}_{\mathrm{clean}} - \overline{\mathrm{LD}}_{\mathrm{corrupt}})$, with means over all 128 prompts

(Zhang & Nanda, 2024). Task accuracy is the fraction of clean prompts with LD > 0: for Pythia-160M 0.79–0.99 (seed 3 is the lowest), for Pythia-410M ≥ 0.98 except seed 4 (0.45, mean clean LD -0.24), and 0.98–1.00 for every GPT-2 seed.

Induction score. 32 sequences of 50 random tokens repeated twice; the score of a head is its mean attention, over the second half, from position t to position $t - 49$.

Signatures and alignment. *Attention*: 24 passages of 64 tokens drawn at random from *Alice's Adventures in Wonderland* (Carroll, 1865); a head's signature is the vector of its attention probabilities p_{ij}, $j < i$, over all passages. *Circuits*: with the gain of the attention normalization folded into W^Q, W^K, W^V and the residual mean projected out (LayerNorm discards it), a head's direct-path circuits on the 128 most frequent tokens of the book are $eM_he^\top$ and $eN_hu^\top$, where the rows of e are the centered, RMS-normalized token embeddings and those of u the unembeddings with the final gain folded in; for Pythia, whose scores depend on the offset δ between positions, we use $M_h(\delta) = W_h^Q R(-\delta)(W_h^K)^\top/\sqrt{d_k}$ for $\delta \in \{0, 1, 2, 4, 8, 16, 32\}$, and the query–key blocks share half of the weight with the value–output block. *Spectra*: the singular values of M_h and N_h relative to the largest (logarithmically) and their traces relative to the Frobenius norm, each feature standardized over the heads of the layer in both runs. Circuits and spectra are computed from the weights alone and are invariant to the per-head gauges (checked numerically to float32 round-off) and to rotations of the residual stream, which act on e, u, M_h and N_h together. For two seeds and each layer we compute the Pearson correlation between the signatures of all pairs of heads and match heads by maximum-weight bipartite matching. With attention signatures the mean correlation of matched heads is 0.65, 0.66, 0.92 and 0.88 (Pythia-160M, 410M, GPT-2 small, medium), against 0.50, 0.57, 0.79 and 0.79 for heads with the same index. The circuit matching agrees with the attention matching for 14%, 9%, 67% and 46% of heads and the spectral matching for 24%, 13%, 41% and 24%, against 8%, 6%, 8% and 6% by chance. The null of the index-matched statistic uses 200 uniformly random relabellings within each layer per pair of seeds.

Statistics. Table 8 reports all statistics, with and without the excluded seeds. The scores sorted within each layer are the maximal invariant of the layer-wise relabelling, but their global rank correlation is not a useful replication statistic: sorting imposes the same within-layer order on both models, and its null, with values reassigned at random across layers, is already 0.72–0.84. Per-layer totals and maxima have a null centred at zero.

	IOI importance						Induction score			
pairs	P-160M 36	(all) 45	P-410M 36	(all) 45	G-S 10	G-M 10	P-160M 45	P-410M 45	G-S 10	G-M 10
index-matched, within layer	0.02 ± 0.08	0.01 ± 0.08	0.02 ± 0.04	0.01 ± 0.05	0.00 ± 0.08	-0.01 ± 0.06	0.00 ± 0.08	0.00 ± 0.06	-0.02 ± 0.09	0.01 ± 0.04
null, 95% quantile	0.14 ± 0.01	0.14 ± 0.01	0.09 ± 0.01	0.09 ± 0.01	0.15 ± 0.02	0.09 ± 0.01	0.14 ± 0.01	0.09 ± 0.01	0.15 ± 0.01	0.09 ± 0.01
aligned by attention patterns	0.24 ± 0.09	0.24 ± 0.08	0.11 ± 0.05	0.10 ± 0.07	0.31 ± 0.10	0.22 ± 0.07	0.60 ± 0.06	0.41 ± 0.08	0.74 ± 0.04	0.61 ± 0.04
aligned by weight circuits	0.04 ± 0.09	0.03 ± 0.09	0.02 ± 0.06	0.01 ± 0.06	0.35 ± 0.05	0.17 ± 0.07	0.12 ± 0.09	0.10 ± 0.07	0.74 ± 0.07	0.57 ± 0.02
aligned by weight spectra	0.16 ± 0.09	0.16 ± 0.09	0.07 ± 0.06	0.06 ± 0.06	0.21 ± 0.10	0.09 ± 0.03	0.44 ± 0.10	0.32 ± 0.10	0.67 ± 0.03	0.44 ± 0.03
top-10 overlap, index-matched	0.12 ± 0.10	0.12 ± 0.09	0.05 ± 0.06	0.05 ± 0.07	0.22 ± 0.10	0.05 ± 0.07	0.15 ± 0.11	0.04 ± 0.07	0.15 ± 0.09	0.08 ± 0.07
top-10 overlap, attention-aligned	0.32 ± 0.14	0.32 ± 0.13	0.14 ± 0.11	0.14 ± 0.10	0.51 ± 0.17	0.24 ± 0.14	0.43 ± 0.16	0.20 ± 0.16	0.49 ± 0.11	0.40 ± 0.15
per-layer totals	0.43 ± 0.27	0.42 ± 0.27	0.17 ± 0.25	0.14 ± 0.24	0.87 ± 0.05	0.36 ± 0.26	0.61 ± 0.18	0.42 ± 0.20	0.76 ± 0.16	0.66 ± 0.11
per-layer maxima	0.61 ± 0.16	0.60 ± 0.15	0.50 ± 0.16	0.47 ± 0.16	0.77 ± 0.08	0.62 ± 0.09	0.59 ± 0.22	0.48 ± 0.23	0.63 ± 0.21	0.61 ± 0.11

Table 8: Agreement between seeds (Spearman ρ or overlap, mean $\pm$ s.d. over pairs of seeds). P: Pythia, G-S and G-M: GPT-2 small and medium. Pythia IOI columns without "(all)" exclude the seeds that do not solve the task; every GPT-2 seed solves it.

D Putting the Criterion to Work

Section 4.4 states when a claim is well posed; this appendix collects how to obtain such claims in practice. A quantity is invariant under the symmetries of Section 4 if it is computed from the function itself (outputs, activations or attention patterns on chosen inputs) or from the invariants of Section 4. For the objects in common use, three operations suffice, and a numerical test checks the result.

Three operations. (i) *Contract the indices on which gauges act.* A gauge acts inside a head, between W_h^Q and W_h^K and between W_h^V and W_h^O, and inside a SwiGLU neuron, between its input

and output weights; products such as M_h, N_h or a neuron's contribution to the output do not see it. (ii) *Read residual directions through the model's own maps or activations.* A change of residual basis acts on embeddings, readers, writers and activations together, so quantities that pair them are invariant: circuits such as $eM_he^\top$ and $eN_hu^\top$ through the model's own embeddings and unembedding, projections of activations on directions defined by its readers or by a probe trained on them, Gram matrices of activations. Coordinates and raw matrix entries are not. (iii) *Treat labels as pointers.* Heads, neurons and experts are permuted by symmetries: a claim may use their labels within a checkpoint, where it is then equivariant, but across runs it must align them by invariant signatures or aggregate over them, and report the index-matched null (Section 5.2). Table 9 applies these operations to common objects.

Object	Changed by	Invariant replacement
Norms or spectra of $W_h^Q, W_h^K, W_h^V, W_h^O$	per-head gauges	the same statistics of M_h and N_h, e.g. $\|N_h\|_F$
Entries of M_h, N_h, readers or writers in residual coordinates	residual rotations (pre-norm)	circuits through the model's embeddings and unembedding; spectra and traces of M_h and N_h
"Head (l, h)" in another run	relabelling	heads matched by attention patterns or circuits (Section 5.2); per-layer statistics
Residual dimension i (outlier, massive activation)	residual rotations (pre-norm)	norms and inner products of residual vectors; projections on directions defined by readers or probes; basis alignment as a property of training, compared across runs
Axes of a subspace read by a probe or a dictionary	change of basis within the subspace	the subspace (its projector $UU^\top$), unless an extra assumption such as sparsity selects axes
Activation of a SwiGLU neuron	rescaling of the neuron	its contribution to the output (activation times output weights)
Index of an expert	expert permutation	router-weighted contribution; experts matched across runs
Scale of the residual stream	global rescaling ($\varepsilon = 0$)	ratios, cosines, normalized quantities

Table 9: Common objects of interpretability, the exact symmetry that changes them, and an invariant replacement.

A numerical test. Whether a quantity is well posed can be checked without any theory: recompute it on weights transformed by random exact symmetries and compare values. A function claim must not change; a checkpoint claim must change by the corresponding relabelling. The released library provides the transformation as one call, `apply_random_symmetries`, which applies, in every layer, head relabellings, query–key and value–output gauges (with the refinements of Section 4.2) and neuron or expert permutations, then a random change of residual basis, and returns the relabellings. On tiny random instances of all seven architecture families it changes the logits by at most 1.3×10^{-14}, and on GPT-2, Pythia-160M and SmolLM2-135M by at most 4.9×10^{-12} (script `run_symmetry_test_demo.py`). On these three models the attention patterns of a middle layer pass as a checkpoint claim (after relabelling they match to 2.4×10^{-12}, whereas by index they differ by 0.7–1.0), $\|N_h\|_F$ passes as a function claim (relative change 10^{-15}), and $\|W_h^O\|_F$ fails (up to 55%). Values must be compared, not only rankings: a random gauge barely reorders the heads by $\|W_h^O\|_F$ (Spearman $\rho \geq 0.99$), whereas an adversarial one reverses the order (Appendix B.8). The test is necessary, not sufficient: it probes the symmetries we know, and a whole network may have others (Section 3).

AI USE STATEMENT

Generative AI tools were used for proofreading and language refinement. The authors assume full responsibility for the final content, including all text, claims, and artifacts generated with AI assistance.

ETHICS STATEMENT

This is a theoretical and methodological study of which parts of a transformer are determined by its function. It involves no human subjects, private data or new dataset. It uses publicly released, ungated pretrained models (listed in Table 1 and Appendices B.5 and C) and a public-domain text from Project Gutenberg, only as inputs to measurements. The paper's message, that interpretability claims

should be stated in terms that the model actually determines, is intended to make interpretability research more reliable; we are not aware of a harmful use of the results.

REPRODUCIBILITY STATEMENT

All proofs are in Appendix A. The anonymized supplementary material contains the code for every experiment and the raw results behind every number and figure: a script that downloads each model at a pinned Hugging Face commit, a library of exact symmetry operations for seven architecture families with a self-test on randomly initialized instances, the atlas, robustness, precision, genericity, reachable-state, weight-statistics and cross-seed scripts, the scripts that generate every table and figure from the result files, and a function that applies random exact symmetries to a checkpoint, with which any interpretability claim can be tested (Appendix D). The models up to 7B parameters run on a CPU in float64; the larger models need a GPU with enough memory to hold them in bfloat16 (we used 128 GB of unified memory).