REPRESENTATION SYMMETRIES AND ARCHITECTURAL CONSTRAINTS ON PROBE IDENTIFIABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

A central question in neural representation analysis is when observations of hidden representations uniquely determine an associated symbolic assignment. A symbolic label may be predicted accurately from a representation without being uniquely determined by the observed information. Architectural information can restrict hidden-coordinate changes that preserve a neural model and its observations, but invariance under known symmetries does not by itself imply identifiability. We distinguish observational identifiability, symmetry invariance, and agreement of fixed linear sign-probe predictions. A general observation-factorization criterion characterizes exact extractability, while full-space and boundary-local probe criteria identify the positive-ray condition for exact agreement. We then study finite representation calibration, where allowed orthogonal transformations fix specified hidden states pointwise. For an explicit norm-and-inner-product observation protocol on a known linear subspace, we classify every observational fibre and characterize both global and fibrewise probe identifiability. Under Gaussian evaluation, the largest disagreement over all calibration-preserving transformations has a closed form determined by the uncalibrated component of the effective probe and is attained by a Householder reflection. Independent Gaussian calibration gives an exact finite-sample law governed by the support dimension. Architectural stabilizers provide upper bounds, not automatic completeness certificates. FP64 experiments support the calibration predictions. On pretrained GPT-2, residual reparameterizations and calibration-preserving attention interventions change fixed-probe predictions, whereas covariantly transformed probes retain the original predictions and complete-model outputs agree to numerical precision.

1 INTRODUCTION

Neural representation analysis asks whether learned latent states carry interpretable symbolic structure. Probing studies linguistic and structural information in hidden representations (Hewitt & Manning, 2019; Tenney et al., 2019; Hewitt & Liang, 2019); network dissection and concept bottlenecks connect internal states to interpretable attributes (Bau et al., 2017; Koh et al., 2020); causal abstraction and neuro-symbolic learning relate neural states to causal or symbolic structure (Geiger et al., 2021; 2022; van Bekkum et al., 2021). Other work studies representation identifiability and model symmetries (Roeder et al., 2021; Marconato et al., 2025; Godfrey et al., 2022; Grigsby et al., 2023). Yet predictability does not imply observational identifiability. A fixed linear sentiment probe may change under a compensated hidden-coordinate reparameterization that preserves model outputs. We therefore ask whether a probe-induced label is determined across all observation-consistent representations.

We study when observations determine a symbolic assignment and how architectural constraints reduce the remaining ambiguity. Observation factorization gives a general identifiability criterion, while observation-preserving transformations provide non-identifiability witnesses. We then specialize the framework to hidden-state reparameterizations and derive full-space and boundary-local criteria for fixed homogeneous sign probes. Architectural constraints further restrict admissible transformations, but generally yield only upper bounds on the actual observation-preserving group rather than classifications of observational fibres. We therefore study finite representation calibra-

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.

tion, where a specified norm-and-inner-product protocol admits exact fibre characterization, probe-identifiability conditions, an attained worst-case Gaussian disagreement, and an exact finite-sample law under independent Gaussian calibration. Seven FP64 experiments test these predictions and architectural constraints; GPT-2 is the main pretrained study, with Pythia (Biderman et al., 2023) providing supplementary evidence.

Relation to prior theory. Roeder et al. (2021) give conditions for representation identifiability up to linear transformations; this does not identify a fixed probe label. Godfrey et al. (2022) characterize ReLU permutation-and-scaling symmetries; Grigsby et al. (2023) study additional parameter symmetries in ReLU networks. We use the known ReLU commutant; Proposition 5 applies the positive-ray criterion to specified groups, rather than classifying network symmetries. Our finite-calibration analysis determines the remaining label ambiguity through explicit fibres and an attained Gaussian disagreement. This specializes task identifiability (Xi & Bloem-Reddy, 2023) to a fixed probe-induced target; it neither establishes learnability nor assigns logical semantics to probe labels.

All proofs and additional technical results are provided in the appendices.

2 Observation-Based Identifiability

Let $\mathfrak{F}$ be a set whose elements $f \in \mathfrak{F}$ are called neural objects. We refer to $\mathfrak{F}$ as the neural representation domain; it may consist of hidden states, neural networks, reparameterized hidden states, transition states, or finite traces.

Let $(\mathfrak{S}, \sim_{\mathfrak{S}})$ be a symbolic domain, where $\mathfrak{S}$ is a set of symbolic objects and $\sim_{\mathfrak{S}}$ is an equivalence relation specifying which objects are regarded as the same identified structure. Examples include programs modulo denotational equivalence, proofs modulo proof equivalence, logical structures modulo isomorphism, and concept labels modulo a specified renaming equivalence.

Definition 1 (Observation protocol and observational equivalence)**.** *An observation protocol on $\mathfrak{F}$ is a finite collection $\mathcal{O} = (\Phi_1, \ldots, \Phi_m)$ of maps $\Phi_i : \mathfrak{F} \to Y_i$. Its joint observation map $\Phi_{\mathcal{O}} : \mathfrak{F} \to Y_{\mathcal{O}}$ is defined by $\Phi_{\mathcal{O}}(f) = (\Phi_1(f), \ldots, \Phi_m(f))$, where $Y_{\mathcal{O}} = \prod_{i=1}^{m} Y_i$.*

For $f, g \in \mathfrak{F}$, define $f \equiv_{\mathcal{O}} g \iff \Phi_{\mathcal{O}}(f) = \Phi_{\mathcal{O}}(g)$. Equivalently, $\Phi_i(f) = \Phi_i(g)$ for every $i = 1, \ldots, m$.

Here each Y_i is the value space of the ith observation. Its elements may represent hidden-state measurements, activation statistics, architectural information, or finite execution traces, depending on specific observation protocols.

A symbolic assignment is a map $\beta : \mathfrak{F} \to \mathfrak{S}$. Here β represents the symbolic structures whose identifiability from observations of neural representations is under study; it is not an algorithm.

Definition 2 (Observation-only extractor and exact extractability)**.** *For a symbolic assignment $\beta : \mathfrak{F} \to \mathfrak{S}$, an observation-only extractor for β is a map $E : \mathrm{im}(\Phi_{\mathcal{O}}) \to \mathfrak{S}/\!\sim_{\mathfrak{S}}$, where $\mathrm{im}(\Phi_{\mathcal{O}})$ is the set of observation values that actually occur. The extractor is correct for β if $E(\Phi_{\mathcal{O}}(f)) = [\beta(f)]_{\sim_{\mathfrak{S}}}$ for every $f \in \mathfrak{F}$. We call β exactly extractable from $\mathcal{O}$ if a correct observation-only extractor for β exists.*

Remark 1 (Scope of exact extractability)**.** *Exact extractability concerns observational identifiability. It does not require the extractor E to be computable, learnable from finite data, or realizable by a chosen probe class.*

Proposition 1 (Exact observation factorization)**.** *Let $\mathcal{O}$ be an observation protocol over $\mathfrak{F}$, and let $\beta : \mathfrak{F} \to \mathfrak{S}$ be a symbolic assignment. Let $q_{\mathfrak{S}} : \mathfrak{S} \to \mathfrak{S}/\!\sim_{\mathfrak{S}}$ be the canonical quotient map defined by $q_{\mathfrak{S}}(s) = [s]_{\sim_{\mathfrak{S}}}$. Let $\pi_{\mathcal{O}} : \mathfrak{F} \to \mathfrak{F}/\!\equiv_{\mathcal{O}}$ be the canonical observational quotient map defined by $\pi_{\mathcal{O}}(f) = [f]_{\equiv_{\mathcal{O}}}$. Then the following statements are equivalent:*

(i) β *is exactly extractable from* $\mathcal{O}$.

(ii) $\forall f, g \in \mathfrak{F},\ f \equiv_{\mathcal{O}} g \Longrightarrow \beta(f) \sim_{\mathfrak{S}} \beta(g)$.

(iii) $\forall y \in \mathrm{im}(\Phi_{\mathcal{O}})$, $q_{\mathfrak{S}} \circ \beta$ *is constant on* $\Phi_{\mathcal{O}}^{-1}(\{y\})$.

(iv) $\exists \overline{\beta}_{\mathcal{O}} : \mathfrak{F}/\!\equiv_{\mathcal{O}} \to \mathfrak{S}/\!\sim_{\mathfrak{S}}$ *such that* $\overline{\beta}_{\mathcal{O}} \circ \pi_{\mathcal{O}} = q_{\mathfrak{S}} \circ \beta$.

Moreover, when these conditions hold, let E denote the unique correct observation-only extractor. Then the map $\widetilde{\Phi}_{\mathcal{O}} : \mathfrak{F}/\equiv_{\mathcal{O}} \to \mathrm{im}(\Phi_{\mathcal{O}})$, defined by $\widetilde{\Phi}_{\mathcal{O}}([f]_{\equiv_{\mathcal{O}}}) = \Phi_{\mathcal{O}}(f)$, is a bijection, and $E = \overline{\beta}_{\mathcal{O}} \circ \widetilde{\Phi}_{\mathcal{O}}^{-1}$.

Definition 3 (Refinement of observation protocols)**.** *Let $\mathcal{O}_1$ and $\mathcal{O}_2$ be observation protocols on the same neural representation domain $\mathfrak{F}$. We say that $\mathcal{O}_2$ refines $\mathcal{O}_1$ if there exists a map $r_{21} : \mathrm{im}(\Phi_{\mathcal{O}_2}) \to \mathrm{im}(\Phi_{\mathcal{O}_1})$ such that $\Phi_{\mathcal{O}_1} = r_{21} \circ \Phi_{\mathcal{O}_2}$.*

Theorem 1 (Monotonicity under observation refinement)**.** *Assume that $\mathcal{O}_2$ refines $\mathcal{O}_1$ through r_{21}. Then:*

(i) for every $f, g \in \mathfrak{F}$, $f \equiv_{\mathcal{O}_2} g \Longrightarrow f \equiv_{\mathcal{O}_1} g$;

(ii) the map $\kappa_{21} : \mathfrak{F}/\equiv_{\mathcal{O}_2} \to \mathfrak{F}/\equiv_{\mathcal{O}_1}$, $\kappa_{21}([f]_{\equiv_{\mathcal{O}_2}}) = [f]_{\equiv_{\mathcal{O}_1}}$, is well defined and surjective;

(iii) if a symbolic assignment $\beta : \mathfrak{F} \to \mathfrak{S}$ is exactly extractable from $\mathcal{O}_1$, then it is exactly extractable from $\mathcal{O}_2$. More precisely, if $E_1 : \mathrm{im}(\Phi_{\mathcal{O}_1}) \to \mathfrak{S}/\sim_{\mathfrak{S}}$ is the correct extractor for $\mathcal{O}_1$, then $E_2 = E_1 \circ r_{21}$ is the correct extractor for $\mathcal{O}_2$.

Remark 2 (Strictness of refinement)**.** *The converse of Theorem 1(iii) need not hold. A refinement may split an observational fibre on which the symbolic assignment is nonconstant, so that the assignment becomes exactly extractable under the refined protocol even though it is not extractable under the coarser one.*

Definition 4 (Non-identifiability witness)**.** *Let $\mathcal{O}$ be an observation protocol over $\mathfrak{F}$, and let $\beta : \mathfrak{F} \to \mathfrak{S}$ be a symbolic assignment. We call $(f, g) \in \mathfrak{F} \times \mathfrak{F}$ a non-identifiability witness for β with respect to $\mathcal{O}$ if $f \equiv_{\mathcal{O}} g$ and $\beta(f) \not\sim_{\mathfrak{S}} \beta(g)$.*

Corollary 1 (Witness characterization of non-identifiability)**.** *A symbolic assignment $\beta : \mathfrak{F} \to \mathfrak{S}$ is not exactly extractable from $\mathcal{O}$ iff there exists a non-identifiability witness for β with respect to $\mathcal{O}$.*

Example 1. *Let $\mathfrak{F} = \{f_0, f_1\}$, let the observation be constant, and let $\ell(f_i) = i$. Even if a probe class contains ℓ, $f_0 \equiv_{\mathcal{O}} f_1$ while $\ell(f_0) \neq \ell(f_1)$. Thus label prediction is not observation-only extraction.*

3 Symmetries of Observation

Some neural-network reparameterizations change internal representations while preserving selected observations, such as rescaling and permuting hidden states without changing the observed outputs. Thus, any symbolic assignment that changes under an observation-preserving reparameterization is not determined by the observations. We model these transformations as group actions because they are composable and invertible on neural objects (Dummit & Foote, 2004).

Let $\mathcal{G}$ be a group acting on $\mathfrak{F}$ from the left, with action written $(g, f) \mapsto g \cdot f$. Thus, for all $g, h \in \mathcal{G}$ and $f \in \mathfrak{F}$, $e \cdot f = f$ and $(gh) \cdot f = g \cdot (h \cdot f)$. The orbit of f is $\mathcal{G} \cdot f = \{g \cdot f : g \in \mathcal{G}\}$.

Definition 5 (Global observation-preserving symmetry)**.** *If $\Phi_{\mathcal{O}}(g \cdot f) = \Phi_{\mathcal{O}}(f)$ for every $f \in \mathfrak{F}$, we call $g \in \mathcal{G}$ globally observation-preserving for $\mathcal{O}$. Define $\mathcal{G}_{\mathcal{O}} = \{g \in \mathcal{G} : \Phi_{\mathcal{O}}(g \cdot f) = \Phi_{\mathcal{O}}(f)$ for every $f \in \mathfrak{F}\}$.*

Definition 6 (Objectwise observation-preserving transformation)**.** *For $f \in \mathfrak{F}$, define $\Gamma_{\mathcal{O}}(f) = \{g \in \mathcal{G} : \Phi_{\mathcal{O}}(g \cdot f) = \Phi_{\mathcal{O}}(f)\}$. An element of $\Gamma_{\mathcal{O}}(f)$ is called observation-preserving at f.*

For every $f \in \mathfrak{F}$, we have $\mathcal{G}_{\mathcal{O}} \subseteq \Gamma_{\mathcal{O}}(f)$, and $\mathcal{G}_{\mathcal{O}} = \bigcap_{f \in \mathfrak{F}} \Gamma_{\mathcal{O}}(f)$. The set $\Gamma_{\mathcal{O}}(f)$ need not be a subgroup and may depend on f. The subgroup $\mathcal{G}_{\mathcal{O}}$ records only transformations that preserve the observations uniformly over the whole domain.

Lemma 1 (Observation-preserving symmetries form a subgroup)**.** *The observation-preserving symmetry set satisfies $\mathcal{G}_{\mathcal{O}} \leq \mathcal{G}$.*

Corollary 2 (Monotonicity of global observation-preserving symmetries)**.** *If $\mathcal{O}_2$ refines $\mathcal{O}_1$ on $\mathfrak{F}$ through r_{21}, then $\mathcal{G}_{\mathcal{O}_2} \leq \mathcal{G}_{\mathcal{O}_1}$.*

A symbolic assignment $\beta : \mathfrak{F} \to \mathfrak{S}$ is invariant under $\mathcal{G}_{\mathcal{O}}$ if $\beta(g \cdot f) \sim_{\mathfrak{S}} \beta(f)$ for every $g \in \mathcal{G}_{\mathcal{O}}$ and $f \in \mathfrak{F}$.

Corollary 3 (Objectwise symmetry obstruction). *Let $\mathcal{G}$ act on $\mathfrak{F}$. If there exist $f \in \mathfrak{F}$ and $g \in \Gamma_{\mathcal{O}}(f)$ such that $\beta(g \cdot f) \not\sim_{\mathfrak{S}} \beta(f)$, then β is not exactly extractable from $\mathcal{O}$.*

Definition 7 (Orbit-separating observation protocol relative to a subgroup). *Let $\mathcal{G}$ act on $\mathfrak{F}$, and let $K \leq \mathcal{G}$ be a subgroup that is globally observation-preserving for $\mathcal{O}$, so that $\Phi_{\mathcal{O}}(k \cdot f) = \Phi_{\mathcal{O}}(f)$ for every $k \in K$ and every $f \in \mathfrak{F}$. We say that $\mathcal{O}$ is orbit-separating relative to K on $\mathfrak{F}$ if, for every $f, h \in \mathfrak{F}$, $f \equiv_{\mathcal{O}} h \iff$ there exists $k \in K$ such that $h = k \cdot f$.*

Corollary 4 (Orbit specialization of exact extractability). *Let $\mathcal{G}$ act on $\mathfrak{F}$, let $K \leq \mathcal{G}_{\mathcal{O}}$, and assume that $\mathcal{O}$ is orbit-separating relative to K on $\mathfrak{F}$. Then a symbolic assignment $\beta : \mathfrak{F} \to \mathfrak{S}$ is exactly extractable from $\mathcal{O}$ iff $\beta(k \cdot f) \sim_{\mathfrak{S}} \beta(f)$ for every $k \in K$ and every $f \in \mathfrak{F}$.*

Proposition 1 gives the exact criterion for arbitrary observation protocols. Because $K \leq \mathcal{G}_{\mathcal{O}}$, every K-orbit is contained in an observational equivalence class. Consequently, exact extractability always implies invariance under K. Orbit separation is a sufficient condition under which this necessary invariance condition is also sufficient for a fixed symbolic assignment. It is not, however, necessary for the exact extractability of a particular assignment. The following proposition gives the precise uniform statement.

Proposition 2 (Orbit separation as a uniform invariance criterion). *Let $\mathcal{G}$ act on $\mathfrak{F}$, let $\mathcal{O}$ be an observation protocol on $\mathfrak{F}$, and let $K \leq \mathcal{G}_{\mathcal{O}}$. Equip $\{0,1\}$ with the equality relation. Then the following statements are equivalent:*

(i) $\mathcal{O}$ is orbit-separating relative to K on $\mathfrak{F}$.

(ii) For every symbolic assignment $\beta : \mathfrak{F} \to \{0,1\}$, the assignment β is exactly extractable from $\mathcal{O}$ iff $\beta(k \cdot f) = \beta(f)$ for every $k \in K$ and every $f \in \mathfrak{F}$.

4 Hidden-State Reparameterization

Fix an input domain $\mathcal{X}$ and integers $d, k \geq 1$. In Euclidean coordinates, we represent covectors by column vectors, so the dual action A^* is represented by $A^\top$.

Definition 8 (Factorized neural mechanism). *A factorized neural mechanism with hidden space $\mathbb{R}^d$ is a pair $\theta = (h, W)$, where $h : \mathcal{X} \to \mathbb{R}^d$ is a hidden representation map and $W : \mathbb{R}^d \to \mathbb{R}^k$ is a linear readout. Let $\Theta_d = \{(h, W) : h : \mathcal{X} \to \mathbb{R}^d, W : \mathbb{R}^d \to \mathbb{R}^k \text{ is linear}\}$. For $\theta = (h, W) \in \Theta_d$, define its input-output map by $y_\theta = W \circ h$.*

Definition 9 (Hidden-state reparameterization action). *For $A \in \mathrm{GL}(d)$ and $\theta = (h, W) \in \Theta_d$, define $A \cdot \theta = A \cdot (h, W) = (A \circ h, W \circ A^{-1})$.*

This is a left action of $\mathrm{GL}(d)$ on Θ_d. Indeed, $I \cdot \theta = \theta$, and for $A, B \in \mathrm{GL}(d)$, $A \cdot (B \cdot (h, W)) = ((AB) \circ h, W \circ (AB)^{-1}) = (AB) \cdot (h, W)$, where I denotes the identity matrix in $\mathrm{GL}(d)$. Moreover, for every $x \in \mathcal{X}$, $y_{A \cdot \theta}(x) = (W \circ A^{-1})(Ah(x)) = Wh(x) = y_\theta(x)$. Let $\Theta_{\mathrm{adm}} \subseteq \Theta_d$ be the chosen admissible mechanism class. In this section, the neural representation domain is $\mathfrak{F} = \Theta_{\mathrm{adm}}$.

Definition 10 (Behavioral observation protocol). *Let $\mathcal{O}_{\mathrm{beh}} = (\Phi_{\mathrm{beh}})$ denote the behavioral observation protocol on Θ_{adm}, where $\Phi_{\mathrm{beh}} : \Theta_{\mathrm{adm}} \to 2^{\mathcal{X} \times \mathbb{R}^k}$ is defined by $\Phi_{\mathrm{beh}}(\theta) = \{(x, y_\theta(x)) : x \in \mathcal{X}\}$.*

Proposition 3 (Behavioral reparameterization witness). *Let $\theta \in \Theta_{\mathrm{adm}}$ and let $A \in \mathrm{GL}(d)$ satisfy $A \cdot \theta \in \Theta_{\mathrm{adm}}$. Then $A \cdot \theta \equiv_{\mathcal{O}_{\mathrm{beh}}} \theta$. If a symbolic assignment $\beta : \Theta_{\mathrm{adm}} \to \mathfrak{S}$ satisfies $\beta(A \cdot \theta) \not\sim_{\mathfrak{S}} \beta(\theta)$ for some pair (A, θ), then β is not exactly extractable from $\mathcal{O}_{\mathrm{beh}}$.*

We use the term gauge (Cohen et al., 2019) to describe the freedom among reparameterizations that preserve a chosen observation protocol. If a protocol observes additional data such as hidden coordinates, readout covectors (Axler, 2024), a metric, or other architectural data, then only reparameterizations that still preserve those observed data remain gauges for that protocol.

Example 2 (Linear probes and gauge dependence). *Let $h \in \mathbb{R}^d$, and let $w \in \mathbb{R}^d \setminus \{0\}$ be a fixed-basis probe direction. Define the linear sign probe by $p_w(h) = \mathrm{sign}(w^\top h) \in \{-1, 0, 1\}$. If $h' = Ah$, the transformed probe vector $w' = A^{-\top} w$ preserves this label since $(w')^\top h' = (A^{-\top} w)^\top Ah = w^\top h$. This joint transformation guarantees label preservation; keeping w fixed need not preserve the label on h and Ah.*

Remark 3 (Probe-induced targets and probe parameters)**.** *The object studied below is the probe-induced target $\beta_w(h) = \operatorname{sign}(w^\top h)$, not the parameter vector w itself. On $\mathbb{R}^d$, two nonzero probe vectors induce the same sign target only if they differ by a positive scalar. The target therefore determines at most the positive ray $\mathbb{R}_{>0}w$, not a unique representative. On a proper support $\mathcal{H} \subsetneq \mathbb{R}^d$, even probe vectors from different positive rays may induce the same target. Accordingly, the extractability statements below concern β_w unless a parameter-valued symbolic assignment is introduced explicitly.*

Let $\mathcal{G} \leq \mathrm{GL}(d)$ be the gauge group representing the group of allowed hidden-state reparameterizations that preserve the observations. We now specialize the neural representation domain to a nonempty $\mathcal{G}$-invariant set of hidden states $\mathcal{H} \subseteq \mathbb{R}^d$, that is, $A(\mathcal{H}) = \mathcal{H}$ for every $A \in \mathcal{G}$. We then characterize exactly when the labels produced by a fixed linear probe are exactly extractable under the remaining representation transformations.

For $h \in \mathcal{H}$, its gauge orbit under $\mathcal{G}$ is $\{Ah : A \in \mathcal{G}\}$. We write $h \equiv_{\mathcal{G}} h' \iff \exists A \in \mathcal{G}$ such that $h' = Ah$.

Theorem 2 (Fixed homogeneous sign-probe gauge criterion)**.** *Let $\mathcal{G} \leq \mathrm{GL}(d)$ act on a nonempty $\mathcal{G}$-invariant set $\mathcal{H} \subseteq \mathbb{R}^d$ by $h \mapsto Ah$. Let $\mathcal{O}$ be an observation protocol on $\mathcal{H}$. Fix $w \in \mathbb{R}^d \setminus \{0\}$, equip the label space $\{-1, 0, 1\}$ with equality, and define $\beta_w : \mathcal{H} \to \{-1, 0, 1\}$ by $\beta_w(h) = \operatorname{sign}(w^\top h)$.*

Assume that, $\forall h, h' \in \mathcal{H}$, one has $h \equiv_{\mathcal{O}} h'$ iff $\exists A \in \mathcal{G}$ such that $h' = Ah$. Then β_w is exactly extractable from $\mathcal{O}$ iff $\operatorname{sign}(w^\top Ah) = \operatorname{sign}(w^\top h)$ for every $A \in \mathcal{G}$ and every $h \in \mathcal{H}$. When these equivalent conditions hold, every $\mathcal{G}$-orbit is contained in one of the three sign fibres $\{h \in \mathcal{H} : w^\top h > 0\}$, $\{h \in \mathcal{H} : w^\top h = 0\}$, or $\{h \in \mathcal{H} : w^\top h < 0\}$.

If $\mathcal{H} = \mathbb{R}^d$, the preceding condition is equivalent to the requirement that, for every $A \in \mathcal{G}$, there exists a unique scalar $\chi(A) > 0$ such that $A^\top w = \chi(A)w$.

On the full hidden space, this means that a nonzero fixed-basis probe is identifiable only when every allowed reparameterization preserves its probe direction up to positive scale.

Corollary 5 (Full-$\mathrm{GL}(d)$ fixed homogeneous sign-probe obstruction)**.** *Let $d \geq 1$, let $w \in \mathbb{R}^d \setminus \{0\}$, and let $\mathcal{O}$ be an observation protocol on $\mathbb{R}^d$. Assume $\forall h, h' \in \mathbb{R}^d$, one has $h \equiv_{\mathcal{O}} h'$ iff $h' = Ah$ for some $A \in \mathrm{GL}(d)$. Then the fixed-basis sign-probe target $\beta_w(h) = \operatorname{sign}(w^\top h)$ is not exactly extractable from $\mathcal{O}$.*

Thus, the same gauge orbit contains two observationally indistinguishable hidden states, one assigned label 1 and the other assigned label -1 by the fixed probe.

Theorem 3 (Boundary-local fixed homogeneous sign-probe gauge criterion)**.** *Let $w \in \mathbb{R}^d$ be nonzero and let $\mathcal{H} \subseteq \mathbb{R}^d$. Assume that $\exists h_0 \in \mathcal{H}$ and $\exists \varepsilon > 0$ such that $w^\top h_0 = 0$ and $B^\circ_\varepsilon(h_0) \subseteq \mathcal{H}$, where $B^\circ_\varepsilon(h_0) = \{h \in \mathbb{R}^d : \|h - h_0\|_2 < \varepsilon\}$. For any $A \in \mathrm{GL}(d)$, the following statements are equivalent:*

(i) $\operatorname{sign}(w^\top Ah) = \operatorname{sign}(w^\top h)$ *for every $h \in \mathcal{H}$.*

(ii) There exists a unique scalar $\chi_A \in \mathbb{R}_{>0}$ such that $A^\top w = \chi_A w$.

Consequently, for a gauge group $\mathcal{G} \leq \mathrm{GL}(d)$, $\operatorname{sign}(w^\top Ah) = \operatorname{sign}(w^\top h)$ for every $A \in \mathcal{G}$ and every $h \in \mathcal{H}$ iff there exists a unique map $\chi : \mathcal{G} \to \mathbb{R}_{>0}$ such that $A^\top w = \chi(A)w$ for every $A \in \mathcal{G}$.

Theorem 3 shows that full-space support is not required to derive the positive-ray condition. An open neighborhood of a point on the decision boundary is already sufficient. By contrast, invariance on a low-dimensional reachable set or on a set lying entirely on one side of the boundary establishes only equality of the induced labels on that set. It does not identify the positive ray of the probe covector on the ambient hidden space.

5 Architectural Constraints and Their Scope

Architectural data restrict the transformations considered above, but an upper bound on a symmetry group must not be treated as a classification of observational fibres. Let $\mathrm{GL}(H)$ act on an ambient domain $\mathfrak{F}$, let $\mathfrak{F}_{\mathrm{adm}}$ be its nonempty admissible subclass, and let $\mathcal{O}$ be an observation protocol on

that subclass. The actual global linear symmetry group consists of those A that map $\mathfrak{F}_{\mathrm{adm}}$ onto itself and satisfy $\Phi_{\mathcal{O}}(A \cdot f) = \Phi_{\mathcal{O}}(f)$ for every admissible f.

Suppose each architectural datum takes values in a space with a specified left group action, and its data map is equivariant, $\tau_j(A \cdot f) = A \cdot \tau_j(f)$. If $\tau_j(f) = T_j$ throughout the admissible class, then

$$\mathcal{G}_{\mathrm{act}} \leq \bigcap_j \mathrm{Stab}_{\mathrm{GL}(H)}(T_j),\ \mathrm{Stab}_{\mathrm{GL}(H)}(T_j) = \{A : A \cdot T_j = T_j\}. \tag{1}$$

Indeed, class preservation gives $\tau_j(A \cdot f) = T_j$, while equivariance gives $\tau_j(A \cdot f) = A \cdot T_j$. If the data vary with f but are recoverable from $\Phi_{\mathcal{O}}(f)$, the analogous bound intersects the stabilizers of every attained value $\tau_j(f)$.

The natural actions yield familiar constraints. A positive definite metric C requires $A^\top C A = C$, a fixed readout U requires $UA = U$, and named block projections P_b require $AP_b = P_b A$. A fixed coordinatewise ReLU requires $\sigma A = A\sigma$ on its full coordinate domain. Its invertible linear commutant is the positive monomial group, the known permutation-and-positive-scaling result of Godfrey et al. (2022). These equations restrict global linear reparameterizations; finite-sample commutation or a change in an unconstrained parameterization does not establish them globally.

Equality in Equation 1 requires independently checking that every candidate transformation preserves both the admissible class and the complete observation map. The corresponding equivalences in Appendix A record these two requirements; their completeness conditions are not an architecture-derived certificate. Likewise, the exchangeable-block semidirect product there represents already specified blockwise actions and permutations, rather than classifying all equivalent neural networks.

For a fixed probe, preservation of its positive ray by an upper-bound group is sufficient for invariance under the actual subgroup. The converse implication cannot be used for an obstruction: a transformation in an upper bound that changes the probe must also be shown to be an admissible observation-preserving transformation. Even an exact global group classification does not establish that all observational fibres are its orbits. The explicit protocol in Section 6 satisfies that stronger condition by a direct argument. Full definitions, block formulas, and their proofs are retained in Appendix A and the subsequent proof sections.

6 Fixed Probes under Finite Representation Calibration

The preceding obstruction is qualitative. We now ask how much ambiguity remains when a transformation must fix a finite collection of hidden states, not merely their probe labels. We therefore consider a concrete observation protocol with classifiable fibres and an exact worst-case disagreement calculation. Here calibration means fixing representation vectors; it does not refer to probability calibration.

Fix a known nonzero subspace $L \leq \mathbb{R}^d$ of dimension r_L, and calibration states $b_1, \ldots, b_{n_{\mathrm{cal}}} \in L$. Let $B_{\mathrm{cal}} = [b_1\ \cdots\ b_{n_{\mathrm{cal}}}]$, let $S_{\mathrm{cal}} = \mathrm{span}\{b_1, \ldots, b_{n_{\mathrm{cal}}}\}$, and write P_L and P_{cal} for the orthogonal projections onto these two subspaces. For $n_{\mathrm{cal}} = 0$, use the empty matrix and $S_{\mathrm{cal}} = \{0\}$. Define

$$\begin{aligned} \Phi_{\mathrm{cal}}(h) &= (B_{\mathrm{cal}}^\top h, \|h\|_2^2),\ h \in L, \\ \mathcal{K}_{\mathrm{cal}} &= \{A \in O(d) : A|_{L^\perp} = I_{L^\perp},\ Ab_i = b_i \text{ for every } i\}. \end{aligned} \tag{2}$$

The calibration states are fixed protocol data, shared across every object being compared. For a fixed probe w, put $u_w = P_L w$ and $v_{\mathrm{cal}} = (P_L - P_{\mathrm{cal}})w$. Assume $u_w \neq 0$; otherwise β_w is identically zero on L.

Theorem 4 (Exact fibres and fixed-probe identifiability)**.** *For any calibration matrix, including rank-deficient ones, the fibres of Φ_{cal} on L are exactly the $\mathcal{K}_{\mathrm{cal}}$-orbits. Moreover,*

$$\mathcal{K}_{\mathrm{cal}} \simeq O(r_L - \mathrm{rank}\, B_{\mathrm{cal}}),\ \beta_w \text{ is identifiable from } \Phi_{\mathrm{cal}} \iff u_w \in S_{\mathrm{cal}}. \tag{3}$$

Within $\mathrm{GL}(L)$, the global observation-preserving group consists precisely of the restrictions of $\mathcal{K}_{\mathrm{cal}}$ to L. For a single observed fibre through h, define $a_h = w^\top P_{\mathrm{cal}} h,\ R_h = \|v_{\mathrm{cal}}\|_2\, \|(P_L - P_{\mathrm{cal}})h\|_2$. The minimum and maximum probe scores on that fibre are $a_h - R_h$ and $a_h + R_h$. Its ternary sign label is unique if and only if $R_h = 0$ or $|a_h| > R_h$.

The result concerns a restricted orthogonal model rather than a classification of equivalent Transformers. Its orbit statement follows by extending an isometry on the unmeasured subspace. Unlike an assumed stabilizer-completeness condition, the observation map and its full linear symmetry group are both explicit. The strict inequality in the last condition is necessary because $\operatorname{sign}(0) = 0$ is a distinct label.

Theorem 5 (Attained disagreement after calibration)**.** *Let $h \sim \mathcal{N}(0, P_L)$ be an independent evaluation state and define $p_{L,w}(A) = \Pr\{\operatorname{sign}(w^\top Ah) \neq \operatorname{sign}(w^\top h)\}$, $\eta_{\mathrm{cal}} = \frac{\|v_{\mathrm{cal}}\|_2}{\|u_w\|_2}$. For every fixed calibration matrix,*

$$\Delta_{\mathrm{cal}}(w) := \max_{A \in \mathcal{K}_{\mathrm{cal}}} p_{L,w}(A) = \frac{2}{\pi} \arcsin \eta_{\mathrm{cal}}. \tag{4}$$

When $v_{\mathrm{cal}} \neq 0$, the maximum is attained by the Householder reflection

$$A_{\mathrm{ref}} = I - 2\frac{v_{\mathrm{cal}} v_{\mathrm{cal}}^\top}{\|v_{\mathrm{cal}}\|_2^2}. \tag{5}$$

It fixes every calibration state, has condition number one, and satisfies $\operatorname{rank}(A_{\mathrm{ref}} - I) = 1$. When $v_{\mathrm{cal}} = 0$, all allowed transformations have zero disagreement.

Corollary 6 (Independent Gaussian calibration)**.** *Fix L and w before drawing $b_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, P_L)$. If $0 < n_{\mathrm{cal}} < r_L$, then*

$$\eta_{\mathrm{cal}}^2 \sim \mathrm{Beta}\left(\frac{r_L - n_{\mathrm{cal}}}{2}, \frac{n_{\mathrm{cal}}}{2}\right). \tag{6}$$

Writing $F_{a,b}$ for the $\mathrm{Beta}(a, b)$ cumulative distribution function, for $0 \leq \epsilon \leq 1$,

$$\Pr\{\Delta_{\mathrm{cal}}(w) \leq \epsilon\} = F_{(r_L - n_{\mathrm{cal}})/2,\, n_{\mathrm{cal}}/2}\left(\sin^2 \frac{\pi\epsilon}{2}\right). \tag{7}$$

At $n_{\mathrm{cal}} = 0$, $\Delta_{\mathrm{cal}} = 1$; at $n_{\mathrm{cal}} \geq r_L$, $\Delta_{\mathrm{cal}} = 0$ almost surely. Exact identification therefore requires $n_{\mathrm{cal}} \geq r_L$ almost surely for a fixed nonzero effective probe, whereas approximate agreement follows the displayed finite-sample law.

The Gaussian angle–disagreement identity and the distribution of a random subspace projection are standard ingredients. Our specialization optimizes over all transformations fixing the observed calibration states and supplies an attaining reparameterization. It differs from uniform sign-measurement guarantees (Plan & Vershynin, 2014): calibration constrains the transformation pointwise, rather than observing random binary measurements of two directions. Equation 4 remains valid conditional on any calibration matrix; the beta law additionally requires its stated independence and Gaussian assumptions. A sampled principal subspace is not a certificate of exact support, and neither probability formula is asserted for pretrained language-model features. In exact arithmetic, a compensated value-space reparameterization with A_{ref} supplies a witness fixing the selected calibration contexts while preserving the model output. The complete proofs are in Appendix C.

d

7 Experiments

Experiments used FP64 on an NVIDIA RTX 4090, with CPU support where needed. Controlled constructions tested the theoretical results, while GPT-2 provided the main pretrained interventions. Ten seeds varied constructions, reparameterizations, or calibration selection, with pretrained checkpoints and probes fixed. Uncertainties are sample standard deviations across seed-level means. Appendix F documents all seven experiments.

Observation refinement. Across 960 constructed matrix pairs in dimensions 32, 64, and 128, equal readouts and squared norms coexisted with distinct certified ranks, 5 and 6. The pairs therefore shared coarse observations but belonged to different invertible-linear orbits. Full Gram observations separated every pair. Orthogonal controls preserved the observations with residuals below 2.3×10^{-12}, illustrating the fibre–orbit distinction in Proposition 2. Construction details are in Appendix F.2.

Finite-sample probe tests. For fixed homogeneous sign probes in dimensions 64, 256, and 1024, transformations satisfying $A^\top w = 1.7w$ preserved every sampled prediction, whereas those satisfying $A^\top w = -w$ reversed every sampled prediction. Shears violating the positive-ray condition could nevertheless escape random tests. At shear strength 0.001, none of 4096 Gaussian samples revealed a disagreement in 6 of the 30 dimension–seed configurations; the corresponding count for a ball around a decision-boundary point was 10. The analytic probability of missing every disagreement was 27.14% for either distribution. On the specified codimension-one subspace and strictly one-sided ball, all tested shears preserved all sampled labels, although directed ambient-space and boundary-local witnesses established disagreement. Thus sample agreement must be interpreted relative to the support, rather than as evidence for the ambient-space condition in Theorems 2 and 3.

Architectural constraints. Metric/readout, ReLU, and block controls distinguish admissible symmetries from stabilizer upper bounds (Appendix F.4). At readout rank $d - 1$, a reflection preserved the metric and readout but reversed the remaining-direction probe, despite zero continuous stabilizer dimension. For every tested rank below d, a stabilizer element changed an additional observation. Positive monomial transformations commuted with ReLU, with compensated network errors below 3.6×10^{-15}. A nonmonomial shear commuted on nonnegative states but failed on a mixed-sign witness. Named-block controls fixed block identities, whereas exchangeable-block controls permitted same-type permutations.

Pretrained model and probes. We used the complete pretrained GPT-2 checkpoint with 12 layers and 144 attention heads. Bias-free logistic sentiment probes were fitted at all 12 residual and 144 head locations using 2048 SST-2 (Socher et al., 2013) training examples, with regularization fixed rather than tuned. Residual features were last-token block outputs, and head features were last-token attention-weighted value vectors before output projection. All 156 probes and their shuffled-label controls met the 10^{-9} gradient tolerance. Mean validation accuracies were 79.03% for residual probes and 67.63% for head probes, compared with 49.84% and 49.67% for shuffled-label controls. Cached evaluation used all 872 labeled validation examples; complete-forward verification used 32 fixed validation texts and eight random-token sequences.

Architectural interventions. We tested 20 conditions over ten seeds, covering compensated reparameterizations, calibration-preserving reflections, and counterfactual parameter changes (Appendix F.8). Attention interventions used known query–key and value–output symmetries (Tran et al., 2025; Silva et al., 2025). Residual negation reversed every fixed residual-probe prediction on the verification texts with zero recorded logit error; residual permutation gave $50.08 \pm 2.07\%$ disagreement (Table 1). Transported probes preserved predictions, with intercept correction for affine interventions. Omitting LayerNorm compensation raised the mean relative logit error to 0.930 (Table 7). The orthogonal stabilizer fixing the all-ones vector and commuting with the stored final LayerNorm gain was trivial; this does not classify all model symmetries.

Complete-forward verification. Across GPT-2 compensated interventions and calibration, validation, and random-token inputs, the maximum relative Frobenius logit error was 6.34×10^{-15}. All logit and internal-state checks passed. The $12{,}510$ tensor-interface forward audits recorded no non-FP64 floating tensors; these audits do not inspect arithmetic within device kernels. Parameters were restored exactly after each intervention, and the final reference-text forward reproduced logits and internal states bitwise.

Controlled finite calibration. We tested Theorem 5 in support coordinates with $r_L \in \{16, 32, 64\}$ and ambient-dimension settings $d \in \{64, 256, 1024\}$, using $65{,}536$ independent Gaussian evaluation states per run. Across random, repeated-column, and probe-in-span calibration, the largest absolute difference between the empirical and analytical ten-seed mean reflection disagreements was 1.61×10^{-3}. Probes in the calibration span had zero constrained disagreement, whereas repeated-column controls retained ambiguity. For Corollary 6, $1{,}000$ independently generated calibration matrices per setting gave a largest mean error in η^2_{cal} of 7.30×10^{-3}. Across twelve Kolmogorov–Smirnov tests, the smallest unadjusted p-value was 0.0222; none rejected at the Bonferroni threshold $0.05/12$. Analytic boundary constructions produced reflected scores within 1.66×10^{-11} of zero, illustrating the strict fibrewise condition.

Table 1: GPT-2 compensated interventions on the 32 validation texts. Errors are relative Frobenius logit errors. Fixed-probe disagreement is in percent, with sample standard deviation across ten seed means. Every row had zero logit-argmax disagreement and zero transported-probe disagreement; affine rows use intercept-corrected transport.

Intervention	Mean error	Maximum error	Residual	Head
Identity	0	0	0.00 ± 0.00	0.00 ± 0.00
Query–key compensation	7.34×10^{-16}	3.89×10^{-15}	0.00 ± 0.00	0.00 ± 0.00
Value–output compensation	8.43×10^{-16}	6.02×10^{-15}	0.00 ± 0.00	49.80 ± 0.94
Head permutation	6.96×10^{-16}	4.39×10^{-15}	0.00 ± 0.00	46.08 ± 1.42
MLP permutation	6.31×10^{-16}	6.34×10^{-15}	0.00 ± 0.00	0.00 ± 0.00
Normalization rescaling	6.16×10^{-16}	4.67×10^{-15}	0.00 ± 0.00	0.00 ± 0.00
Residual permutation	1.21×10^{-15}	4.96×10^{-15}	50.08 ± 2.07	0.00 ± 0.00
Residual negation	0	0	100.00 ± 0.00	0.00 ± 0.00
Joint linear transformation	1.30×10^{-15}	4.51×10^{-15}	50.60 ± 1.46	50.07 ± 1.26
Key-bias translation	5.94×10^{-16}	4.09×10^{-15}	0.00 ± 0.00	0.00 ± 0.00
Value-bias translation	7.52×10^{-16}	5.00×10^{-15}	0.00 ± 0.00	34.03 ± 1.26
Residual mean translation	5.83×10^{-16}	4.18×10^{-15}	10.34 ± 7.02	0.00 ± 0.00

Table 2: Calibration-preserving GPT-2 interventions. Head-probe disagreement (%) is reported as mean $\pm$ SD across ten seeds. Cached and complete-forward evaluations use 872 and 32 validation examples, respectively; the final column reports the maximum relative calibration-context error.

n_{cal}	Cached disagreement	Forward disagreement	Max logit error	Max calibration error
8	63.75 ± 1.40	63.97 ± 2.19	4.89×10^{-15}	8.20×10^{-15}
32	40.87 ± 1.14	39.80 ± 1.67	4.20×10^{-15}	8.33×10^{-15}
64	0.00 ± 0.00	0.00 ± 0.00	0	0

Calibration-preserving GPT-2 interventions. Calibration contexts were sampled from a disjoint pool of 256 SST-2 training texts. At each attention head, we constructed a Householder reflection fixing the calibration span, with value–output compensation applied simultaneously to all 144 heads. Table 2 separates cached evaluation on 872 validation examples from complete forwards on 32 verification texts. With eight calibration contexts, cached head-probe disagreement was $63.75 \pm 1.40\%$; with 32, it was $40.87 \pm 1.14\%$. Every selected calibration text was also recomputed through the modified complete model. The largest relative calibration-context error was 8.33×10^{-15}, and all covariantly transformed probes retained their predictions. At 64 calibration contexts, recorded disagreement and forward errors were zero; this finite-sample result does not establish population support.

Discussion. The controlled results agree with the Gaussian calibration theory. GPT-2 interventions show that fixed-probe predictions can change despite preservation of model outputs and selected calibration representations to numerical precision. Covariant probe agreement distinguishes coordinate dependence from information loss. Neither the Gaussian optimum nor the beta law is asserted for pretrained features, and these experiments establish neither population support nor a complete Transformer symmetry classification. Supplementary Pythia results, including optimization diagnostics and complete-forward numerical checks, are reported in Appendices F.5–F.7.

8 Conclusion

This paper separates observational identifiability from invariance under representation symmetries. Full-space and boundary-local criteria characterize exact agreement of fixed homogeneous sign probes. For a finite-calibration protocol, we classify observation fibres and derive fibrewise identification conditions, an attained Gaussian disagreement formula, and an exact law for independent Gaussian calibration. Architectural stabilizers remain upper bounds unless independently shown admissible and observation-preserving. GPT-2 interventions illustrate how fixed-probe predictions depend on representation coordinates, even after finite calibration, while covariant probe transport retains the original predictions.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this paper, to support the retrieval and discovery of related work, and to assist in checking and revising portions of the mathematical arguments and proofs, and to support the verification of portions of the experimental code. All AI-assisted writing was independently verified by the authors against the stated assumptions and derivations, and all references identified with AI assistance were checked against the original publications. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Complete proofs are provided in the appendices. Appendix F documents all seven experiments, including controlled finite calibration in Experiment 5, supplementary Pythia studies in Experiments 4 and 6, and the main GPT-2 study in Experiment 7. Code and configurations are provided in the supplementary material.

REFERENCES

Sheldon Axler. *Linear algebra done right*. Springer, 2024.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network Dissection: Quantifying Interpretability of Deep Visual Representations . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.354. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.354.

Abraham Berman and Robert J Plemmons. *Nonnegative matrices in the mathematical sciences*. SIAM, 1994.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory 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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *International conference on Machine learning*, pp. 1321–1330. PMLR, 2019.

David S Dummit and Richard M Foote. Abstract algebra. john wiley & sons. *Inc., Hoboken, NJ*, 2004.

Atticus Geiger, Hanson Lu, Thomas F Icard, and Christopher Potts. Causal abstractions of neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=RmuXDtjDhG.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338. PMLR, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. *Advances in Neural Information Processing Systems*, 35:11893–11905, 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'23. JMLR.org, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Emanuele Marconato, Sebastien Lachapelle, Sebastian Weichwald, and Luigi Gresele. All or none: Identifiable linear properties of next-token predictors in language modeling. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4123–4131. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/marconato25a.html.

Yaniv Plan and Roman Vershynin. Dimension reduction by random hyperplane tessellations. *Discrete Comput. Geom.*, 51(2):438–461, March 2014. ISSN 0179-5376. doi: 10.1007/s00454-013-9561-6. URL https://doi.org/10.1007/s00454-013-9561-6.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pp. 9030–9039. PMLR, 2021.

Marvin F. Da Silva, Felix Dangel, and Sageev Oore. Hide & seek: Transformer symmetries obscure sharpness & Riemannian geometry finds it. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55591–55613. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/silva25a.html.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170/.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Hoang V. Tran, Thieu Vo, An Nguyen The, Tho Tran Huu, Minh-Khoi Nguyen-Nhat, Thanh Tran, Duy-Tung Pham, and Tan Minh Nguyen. Equivariant neural functional networks for transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uBai0ukstY.

Michael van Bekkum, Maaike de Boer, Frank van Harmelen, André Meyer-Vitali, and Annette ten Teije. Modular design patterns for hybrid learning and reasoning systems: A taxonomy, patterns

and use cases. *Applied Intelligence*, 51(9):6528–6546, 2021. doi: 10.1007/s10489-021-02394-3. URL https://doi.org/10.1007/s10489-021-02394-3.

Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6912–6939. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/xi23a.html.

A Architectural definitions and auxiliary results

This appendix formulates stabilizer bounds for architectural data that are either fixed throughout the admissible class or recoverable from the observation protocol. In both cases, reparameterization covariance constrains the actual global observation-preserving group to the joint stabilizer of the corresponding data values.

Throughout this section, H is a finite-dimensional real vector space, and a basis is fixed only when a coordinate formula is stated. A hidden-state reparameterization is represented by an element $A \in \mathrm{GL}(H)$. A type space on H is a set $\mathcal{D}$ equipped with a left action of $\mathrm{GL}(H)$. A typed architectural datum is a pair $(\mathcal{D}, T)$ with $T \in \mathcal{D}$; when the type space is clear, we refer to T as the datum.

Definition 11 (Natural actions on architectural data)**.** *Let $A \in \mathrm{GL}(H)$, and let $A^* : H^* \to H^*$ be the dual map defined by $(A^*\ell)(h) = \ell(Ah)$. We use the following left actions.*

(i) For $C \in \mathrm{Sym}^2(H^*)$, define $(A \cdot C)(h, h') = C(A^{-1}h, A^{-1}h')$. Then $A \cdot C = C$ iff $C(Ah, Ah') = C(h, h')$ for all $h, h' \in H$. In fixed coordinates, the stabilizer equation is $A^\top C A = C$.

(ii) For $U \in \mathrm{Hom}(H, Y)$, where Y is fixed, define $A \cdot U = U \circ A^{-1}$. Then $A \cdot U = U$ iff $U \circ A = U$.

(iii) For $\ell \in H^*$, define $A \cdot \ell = (A^{-1})^*\ell$. Then $A \cdot \ell = \ell$ iff $A^*\ell = \ell$.

(iv) For $v \in H$, define $A \cdot v = Av$. Its stabilizer equation is $Av = v$.

(v) Let $H = \bigoplus_{b \in B} H_b$ be a named decomposition with canonical projections $\mathbf{P} = (P_b)_{b \in B}$. Define $A \cdot \mathbf{P} = (AP_bA^{-1})_{b \in B}$. Then $A \cdot \mathbf{P} = \mathbf{P}$ iff $AP_b = P_bA$ for every $b \in B$.

(vi) Let $\mathcal{N}(H) = \{(V, \varphi) : V \leq H,\ \varphi : V \to V\}$. Define $A \cdot (V, \varphi) = \big(A(V), A \circ \varphi \circ (A^{-1}|_{A(V)})\big)$. The pair (V, φ) is fixed by A iff $A(V) = V$ and $\varphi \circ (A|_V) = (A|_V) \circ \varphi$.

We use the conventions $H^* = \mathrm{Hom}(H, \mathbb{R})$, $A^*\ell = \ell \circ A$. We identify $\mathrm{Sym}^2(H^*)$ with the space of symmetric bilinear forms on H. The notation $\mathrm{Sym}(B)$ denotes the permutation group of B.

Remark 4 (Coordinate rows of a linear readout)**.** *Let $U : H \to \mathbb{R}^m$, let $e_1^*, \dots, e_m^*$ be the coordinate covectors on $\mathbb{R}^m$, and set $u_j = U^*e_j^* = e_j^* \circ U \in H^*$. Then $U \circ A = U$ iff $A^*u_j = u_j$ for every $j = 1, \dots, m$. After a basis of H is fixed, this becomes $A^\top u_j = u_j$.*

Definition 12 (Joint stabilizer of architectural data)**.** *Let $\mathcal{T} = ((\mathcal{D}_j, T_j))_{j \in J}$ be a family of typed architectural data on H. Define $\mathrm{Stab}_{\mathrm{GL}(H)}(T_j) = \{A \in \mathrm{GL}(H) : A \cdot T_j = T_j\}$ and $\mathcal{G}(\mathcal{T}) = \bigcap_{j \in J} \mathrm{Stab}_{\mathrm{GL}(H)}(T_j)$. When the type spaces are clear, we write $\mathcal{T} = (T_j)_{j \in J}$. Since each type carries a left action, every $\mathrm{Stab}_{\mathrm{GL}(H)}(T_j)$ is a subgroup of $\mathrm{GL}(H)$, and hence $\mathcal{G}(\mathcal{T}) \leq \mathrm{GL}(H)$.*

Assume that $\mathrm{GL}(H)$ acts on an ambient neural representation domain $\mathfrak{F}$. For $A \in \mathrm{GL}(H)$ and $\mathfrak{F}_{\mathrm{adm}} \subseteq \mathfrak{F}$, write $A \cdot \mathfrak{F}_{\mathrm{adm}} = \{A \cdot f : f \in \mathfrak{F}_{\mathrm{adm}}\}$.

Definition 13 (Admissible representation class)**.** *A nonempty subset $\mathfrak{F}_{\mathrm{adm}} \subseteq \mathfrak{F}$ is an admissible representation class if it is the class of neural objects allowed by the architecture, parameterization, observation protocol, and typing assumptions under consideration.*

Definition 14 (Objectwise admissible observation-preserving reparameterization)**.** *Let $\mathcal{O}$ be an observation protocol on $\mathfrak{F}_{\mathrm{adm}}$. For $f \in \mathfrak{F}_{\mathrm{adm}}$, define $\Gamma_{\mathcal{O}}^{\mathrm{adm}}(f) = \{A \in \mathrm{GL}(H) : A \cdot f \in \mathfrak{F}_{\mathrm{adm}}$ and $A \cdot f \equiv_{\mathcal{O}} f\}$.*

The set $\Gamma_{\mathcal{O}}^{\mathrm{adm}}(f)$ may depend on f and need not be a subgroup.

Definition 15 (Global admissible observation-preserving linear gauge group)**.** *For a fixed observation protocol $\mathcal{O}$ on $\mathfrak{F}_{\mathrm{adm}}$, define $\mathcal{G}_{\mathrm{act}} = \left\{ \begin{array}{r} A \in \mathrm{GL}(H) : A \cdot \mathfrak{F}_{\mathrm{adm}} = \mathfrak{F}_{\mathrm{adm}}, \\ \forall f \in \mathfrak{F}_{\mathrm{adm}},\ \Phi_{\mathcal{O}}(A \cdot f) = \Phi_{\mathcal{O}}(f) \end{array} \right\}$. We call $\mathcal{G}_{\mathrm{act}}$ the global admissible observation-preserving linear gauge group of $\mathcal{O}$ on $\mathfrak{F}_{\mathrm{adm}}$. When the dependence must be displayed, we write $\mathcal{G}_{\mathrm{act}}(\mathcal{O}, \mathfrak{F}_{\mathrm{adm}})$.*

Remark 5 (Scope of the global linear gauge group)**.** *The group $\mathcal{G}_{\mathrm{act}}$ captures the invertible linear reparameterizations that preserve the admissible class as a whole and leave all observations unchanged. It does not account for transformations chosen separately for individual neural objects or for nonlinear reparameterizations. Moreover, its orbits may be strictly smaller than the observational fibres of $\mathcal{O}$.*

If $\mathcal{O}_2$ refines $\mathcal{O}_1$ on the same admissible class, then $\mathcal{G}_{\mathrm{act}}(\mathcal{O}_2) \leq \mathcal{G}_{\mathrm{act}}(\mathcal{O}_1)$, where the common admissible class is suppressed. The objectwise and global notions satisfy $A \in \mathcal{G}_{\mathrm{act}} \iff A, A^{-1} \in \bigcap_{f \in \mathfrak{F}_{\mathrm{adm}}} \Gamma_{\mathcal{O}}^{\mathrm{adm}}(f)$. The two-sided condition ensures that A maps the admissible class onto itself, rather than merely into itself.

If $f \in \mathfrak{F}_{\mathrm{adm}}$ and $A \in \Gamma_{\mathcal{O}}^{\mathrm{adm}}(f)$ satisfy $\beta(A \cdot f) \not\sim_{\mathfrak{S}} \beta(f)$, then $(A \cdot f, f)$ is a non-identifiability witness. Hence β is not exactly extractable from $\mathcal{O}$ by Corollary 1.

Definition 16 (Reparameterization-covariant architectural data map)**.** *Let $\mathcal{D}$ be a type space on H. A map $\tau : \mathfrak{F} \to \mathcal{D}$ is reparameterization-covariant if $\tau(A \cdot f) = A \cdot \tau(f)$ for every $A \in \mathrm{GL}(H)$ and every $f \in \mathfrak{F}$. Relative to an admissible class $\mathfrak{F}_{\mathrm{adm}} \subseteq \mathfrak{F}$, we say that τ is fixed with value $T \in \mathcal{D}$ if $\tau(f) = T$ for every $f \in \mathfrak{F}_{\mathrm{adm}}$. We say that τ is observed by an observation protocol $\mathcal{O}$ on $\mathfrak{F}_{\mathrm{adm}}$ if there exists a map $\widehat{\tau} : \mathrm{im}(\Phi_{\mathcal{O}}) \to \mathcal{D}$ such that $\tau|_{\mathfrak{F}_{\mathrm{adm}}} = \widehat{\tau} \circ \Phi_{\mathcal{O}}$.*

Definition 17 (Architecture-defining data fixed on an admissible class)**.** *Let $\boldsymbol{\tau} = (\tau_j)_{j \in J}$ be a family of reparameterization-covariant architectural data maps, where $\tau_j : \mathfrak{F} \to \mathcal{D}_j$. Let $\mathcal{T} = ((\mathcal{D}_j, T_j))_{j \in J}$ be a family of typed architectural data. We say that $\mathcal{T}$ is fixed on $\mathfrak{F}_{\mathrm{adm}}$ through $\boldsymbol{\tau}$ if $\tau_j(f) = T_j$ for every $f \in \mathfrak{F}_{\mathrm{adm}}$ and every $j \in J$.*

The maps in $\boldsymbol{\tau}$ are part of the specification of the neural objects and are fixed before the observation-preserving gauge group is determined. The definition therefore does not assume any inclusion between the gauge group and the joint stabilizer. Instead, that inclusion will follow from reparameterization covariance and preservation of the admissible class.

We say that $\mathfrak{F}_{\mathrm{adm}}$ is invariant under the joint stabilizer $\mathcal{G}(\mathcal{T})$ if $A \cdot \mathfrak{F}_{\mathrm{adm}} = \mathfrak{F}_{\mathrm{adm}}$ for every $A \in \mathcal{G}(\mathcal{T})$. The following completeness definition assumes observation preservation for class-preserving candidates. Together with class invariance, it restates the reverse inclusion needed for equality; it does not provide an independent structural completeness test.

Definition 18 (Completeness of a stabilizer bound)**.** *Let $\mathcal{T}$ be architecture-defining data fixed on $\mathfrak{F}_{\mathrm{adm}}$. We say that the stabilizer bound induced by $\mathcal{T}$ is complete for $\mathcal{O}$ on $\mathfrak{F}_{\mathrm{adm}}$ if every $A \in \mathcal{G}(\mathcal{T})$ satisfying $A \cdot \mathfrak{F}_{\mathrm{adm}} = \mathfrak{F}_{\mathrm{adm}}$ also satisfies $\Phi_{\mathcal{O}}(A \cdot f) = \Phi_{\mathcal{O}}(f)$ for every $f \in \mathfrak{F}_{\mathrm{adm}}$.*

Theorem 6 (Fixed-data architectural stabilizer bound)**.** *Let $\mathfrak{F}_{\mathrm{adm}} \subseteq \mathfrak{F}$ be a nonempty admissible representation class, let $\mathcal{O}$ be an observation protocol on $\mathfrak{F}_{\mathrm{adm}}$, and let $\boldsymbol{\tau} = (\tau_j)_{j \in J}$ be a family of reparameterization-covariant architectural data maps. Assume that $\mathcal{T} = ((\mathcal{D}_j, T_j))_{j \in J}$ is fixed on $\mathfrak{F}_{\mathrm{adm}}$ through $\boldsymbol{\tau}$.*

Let $\mathcal{G}_{\mathrm{act}} = \mathcal{G}_{\mathrm{act}}(\mathcal{O}, \mathfrak{F}_{\mathrm{adm}})$. Then the following statements hold.

(i) $\mathcal{G}_{\mathrm{act}} \leq \mathcal{G}(\mathcal{T}) \leq \mathrm{GL}(H)$.

(ii) $\mathcal{G}_{\mathrm{act}} = \mathcal{G}(\mathcal{T})$ *iff both of the following conditions hold:*

- *(a)* $\mathfrak{F}_{\mathrm{adm}}$ *is invariant under* $\mathcal{G}(\mathcal{T})$*;*
- *(b) the stabilizer bound induced by* $\mathcal{T}$ *is complete for* $\mathcal{O}$ *on* $\mathfrak{F}_{\mathrm{adm}}$.

Exact extractability requires invariance under $\mathcal{G}_{\mathrm{act}}$. If $\mathcal{G}(\mathcal{T})$ preserves the positive ray of a probe covector, this condition holds; under orbit separation relative to $\mathcal{G}_{\mathrm{act}}$, it is also sufficient.

We now turn to architectural data whose values may vary across admissible neural objects. In this setting, preservation of the admissible class alone does not yield a single fixed stabilizer. We therefore require each architectural data map to factor through the observation map and intersect the stabilizers of all values attained on the admissible class.

Let $\boldsymbol{\tau} = (\tau_j)_{j\in J}$ be a family of reparameterization-covariant architectural data maps with type spaces $\mathcal{D}_j$. Define the joint stabilizer of attained data values $\mathcal{G}(\boldsymbol{\tau};\mathfrak{F}_{\mathrm{adm}}) = \bigcap_{f\in\mathfrak{F}_{\mathrm{adm}}}\bigcap_{j\in J}\mathrm{Stab}_{\mathrm{GL}(H)}(\tau_j(f))$.

Definition 19 (Completeness for observed-data stabilizers)**.** *The observed-data constraints induced by $\boldsymbol{\tau}$ are complete for $\mathcal{O}$ on $\mathfrak{F}_{\mathrm{adm}}$ if every $A \in \mathcal{G}(\boldsymbol{\tau};\mathfrak{F}_{\mathrm{adm}})$ satisfying $A\cdot\mathfrak{F}_{\mathrm{adm}} = \mathfrak{F}_{\mathrm{adm}}$ also satisfies $\Phi_{\mathcal{O}}(A\cdot f) = \Phi_{\mathcal{O}}(f)$ for every $f \in \mathfrak{F}_{\mathrm{adm}}$.*

Proposition 4 (Stabilizer bound for model-dependent architectural data)**.** *Assume that every τ_j is observed by $\mathcal{O}$ on $\mathfrak{F}_{\mathrm{adm}}$. Then $\mathcal{G}_{\mathrm{act}} \le \mathcal{G}(\boldsymbol{\tau};\mathfrak{F}_{\mathrm{adm}})$. Moreover, $\mathcal{G}_{\mathrm{act}} = \mathcal{G}(\boldsymbol{\tau};\mathfrak{F}_{\mathrm{adm}})$ iff $\mathfrak{F}_{\mathrm{adm}}$ is invariant under $\mathcal{G}(\boldsymbol{\tau};\mathfrak{F}_{\mathrm{adm}})$ and the observed-data constraints are complete for $\mathcal{O}$ on $\mathfrak{F}_{\mathrm{adm}}$.*

Remark 6 (Fixed and model-dependent architectural data)**.** *Theorem 6 uses data fixed throughout the admissible class; Proposition 4 uses data recoverable from observations. When $\tau_j(f) = T_j$ for every admissible f and every j, the latter stabilizer reduces to $\mathcal{G}(\boldsymbol{\tau};\mathfrak{F}_{\mathrm{adm}}) = \mathcal{G}((T_j)_{j\in J})$.*

Corollary 7 (Metric and readout stabilizer bound)**.** *Let $H = \mathbb{R}^d$, and let $\mathcal{O}$ be an observation protocol on $\mathfrak{F}_{\mathrm{adm}}$. Let $C \in \mathbb{R}^{d\times d}$ be symmetric positive definite and represent the bilinear form $C(h,h') = h^\top Ch'$. Let $U : H \to \mathbb{R}^m$ be a linear readout map. Let $e_1^*,\dots,e_m^*$ be the coordinate covectors on $\mathbb{R}^m$, and define $u_j = U^*(e_j^*) = e_j^*\circ U \in H^*$ for $j \in \{1,\dots,m\}$. Assume that there exist reparameterization-covariant architecture-defining data maps $\tau_C : \mathfrak{F} \to \mathrm{Sym}^2(H^*)$ and $\tau_U : \mathfrak{F} \to \mathrm{Hom}(H,\mathbb{R}^m)$ such that $\tau_C(f) = C$ and $\tau_U(f) = U$ for every $f \in \mathfrak{F}_{\mathrm{adm}}$. Set $\mathcal{T} = \{C,U\}$. Let $\mathcal{G}_{\mathrm{act}}$ be the global admissible observation-preserving linear gauge group of $\mathcal{O}$ on $\mathfrak{F}_{\mathrm{adm}}$. Then $\mathcal{G}_{\mathrm{act}} \le \mathcal{G}(C,U) = \big\{A \in \mathrm{GL}(d) : A^\top CA = C,\ A^\top u_j = u_j \text{ for every } j = 1,\dots,m\big\}$.*

In the variant in which the exposed family is $\mathcal{T} = \{C\}$, the upper-bound group reduces to $O(C) = \{A \in \mathrm{GL}(d) : A^\top CA = C\}$. When $C = I$, this is the ordinary orthogonal group $O(d)$.

Corollary 7 shows that $\mathcal{G}_{\mathrm{act}}$, the group of transformations left unresolved by the full observation protocol, is a subgroup of $\mathcal{G}(C,U)$, the group of transformations that preserve the exposed metric and readout. The converse need not hold. To determine how a coordinatewise ReLU map constrains hidden-state reparameterizations, we characterize all invertible linear maps that commute with it.

Theorem 7 (Known ReLU commutant, Godfrey et al., 2022)**.** *Let $\sigma : \mathbb{R}^d \to \mathbb{R}^d$ be the coordinatewise ReLU map, $\sigma(z)_i = \max\{z_i, 0\}$, for every $z \in \mathbb{R}^d$ and every $i = 1,\dots,d$. Then $\{A \in \mathrm{GL}(d) : \sigma(Az) = A\sigma(z) \text{ for every } z \in \mathbb{R}^d\} = \mathsf{Mon}_{>0}(d)$, where $\mathsf{Mon}_{>0}(d) = \{D\Pi_\pi : D = \mathrm{diag}(\lambda_1,\dots,\lambda_d),\ \lambda_i > 0,\ \pi \in S_d\}$.*

We use S_d to denote the symmetric group on $\{1,\dots,d\}$ and Π_π to denote the permutation matrix associated with $\pi \in S_d$ (Dummit & Foote, 2004). We use the convention $\Pi_\pi e_j = e_{\pi(j)}$, where e_j is the jth standard basis vector of $\mathbb{R}^d$. The positive diagonal matrix D represents positive coordinate rescalings. Thus, $\mathsf{Mon}_{>0}(d)$ is the positive monomial matrix group and is isomorphic to $\mathbb{R}^d_{>0} \rtimes S_d$ (Dummit & Foote, 2004; Berman & Plemmons, 1994).

Remark 7 (Scope of the ReLU commutant)**.** *Theorem 7 classifies the invertible linear maps A satisfying $\sigma\circ A = A\circ\sigma$ on all of $\mathbb{R}^d$ for one fixed, bias-free, coordinatewise ReLU map. It does not classify the complete parameter-symmetry group of a multilayer ReLU network. Biases, normalization layers, gating mechanisms, residual couplings, parameter tying, and architecture-specific admissibility constraints require additional stabilizer equations, and the resulting network-level gauge need not coincide with $\mathsf{Mon}_{>0}(d)$. The positive-monomial conclusion also need not follow when the commutation identity is required only on a finite sample, on a low-dimensional reachable set, or on a proper subset of the ReLU coordinate domain.*

For a linear subspace $V \le \mathbb{R}^d$, write $O(V) = \{Q \in \mathrm{GL}(V) : \langle Qv, Qv'\rangle_2 = \langle v, v'\rangle_2 \text{ for every } v, v' \in V\}$, where $\langle\cdot,\cdot\rangle_2$ is the restriction to V of the standard Euclidean inner product on $\mathbb{R}^d$.

Proposition 5 (Fixed homogeneous sign-probe targets under reduced gauges). *Let $H = \mathbb{R}^d$, let $\mathcal{H} \subseteq H$ be a nonempty admissible hidden-state support, and set $\mathcal{H}$ as the neural representation domain. Let $\mathcal{O}$ be an observation protocol on $\mathcal{H}$, and let $\mathcal{G}_{\mathrm{act}} \leq \mathrm{GL}(H)$ satisfy $A(\mathcal{H}) = \mathcal{H}$ and $\Phi_{\mathcal{O}}(Ah) = \Phi_{\mathcal{O}}(h)$ for every $A \in \mathcal{G}_{\mathrm{act}}$ and every $h \in \mathcal{H}$. Let $w \in H^* \setminus \{0\}$, and define $\beta_w : \mathcal{H} \to \{-1, 0, 1\}$ by $\beta_w(h) = \mathrm{sign}(w(h))$, where the probe labels are compared by equality. Assume that either $\mathcal{H} = H$, or there exist $h_0 \in \mathcal{H}$ and $\varepsilon > 0$ such that $w(h_0) = 0$ and $B^\circ_\varepsilon(h_0) \subseteq \mathcal{H}$.*

If β_w is exactly extractable from $\mathcal{O}$, then for every $A \in \mathcal{G}_{\mathrm{act}}$ there exists a unique scalar $\chi_w(A) > 0$ such that $A^\top w = \chi_w(A) w$. Moreover, $\chi_w : \mathcal{G}_{\mathrm{act}} \to \mathbb{R}_{>0}$ is a group homomorphism. Consequently, the following statements hold.

(i) Assume $H = V \oplus V_{\mathrm{c}}$, $w|_{V_{\mathrm{c}}} = 0$, and $w|_V \neq 0$. If $\{Q \oplus I_{V_{\mathrm{c}}} : Q \in O(V)\} \leq \mathcal{G}_{\mathrm{act}}$, then β_w is not exactly extractable from $\mathcal{O}$.

(ii) Assume $H = H_b \oplus H_{\neg b}$, $w|_{H_{\neg b}} = 0$, and $w|_{H_b} \neq 0$. If $\{B \oplus I_{H_{\neg b}} : B \in \mathrm{GL}(H_b)\} \leq \mathcal{G}_{\mathrm{act}}$, then β_w is not exactly extractable from $\mathcal{O}$.

(iii) Assume $H = H_b \oplus H_{\neg b}$ is a fixed coordinate decomposition. Let $P_b \in \mathrm{End}(H)$ be the canonical projection with $\mathrm{im}(P_b) = H_b$ and $\ker(P_b) = H_{\neg b}$. Assume that $AP_b = P_b A$ for every $A \in \mathcal{G}_{\mathrm{act}}$. Fix a coordinate identification $H_b \simeq \mathbb{R}^{d_b}$, and define the induced block gauge by $\mathcal{G}_b = \{A|_{H_b} : A \in \mathcal{G}_{\mathrm{act}}\}$. Assume that $\mathcal{G}_b \leq \mathsf{Mon}_{>0}(d_b)$. Let $e_i^* \in H^*$ be the zero extension to H of the i th coordinate covector on H_b, and suppose $w = e_i^*$. Every matrix in $\mathsf{Mon}_{>0}(d_b)$ has a unique factorization $M = D\Pi_\pi$ with D positive diagonal and $\pi \in S_{d_b}$. Define $\mathrm{perm}_b : \mathsf{Mon}_{>0}(d_b) \to S_{d_b}$ by $\mathrm{perm}_b(D\Pi_\pi) = \pi$. The uniqueness of the factorization and multiplication of the permutation factors show that perm_b is a group homomorphism. Define $\mathfrak{P}_b = \mathrm{perm}_b(\mathcal{G}_b)$. Then $\mathfrak{P}_b \leq S_{d_b}$.

If $\beta_{e_i^*}$ is exactly extractable from $\mathcal{O}$, then $\pi(i) = i$ for every $\pi \in \mathfrak{P}_b$. If, in addition, $\mathcal{O}$ is orbit-separating relative to $\mathcal{G}_{\mathrm{act}}$ on $\mathcal{H}$, then $\beta_{e_i^*}$ is exactly extractable from $\mathcal{O}$ if and only if $\pi(i) = i$ for every $\pi \in \mathfrak{P}_b$.

This proposition recovers Corollary 5 as the special case $\mathcal{G}_{\mathrm{act}} = \mathrm{GL}(d)$. The full general linear gauge contains transformations that do not preserve the positive ray of any fixed nonzero probe covector. Smaller architectural gauge groups may permit exact extraction of a probe-induced target only when the positive ray of its defining covector is invariant under the remaining group. Orthogonal, blockwise general linear, and positive monomial gauges differ in the covector rays they preserve.

The following named-block construction is abstract. Attention-head value spaces are not assumed to be a direct-sum decomposition of an actual Transformer residual stream.

Definition 20 (Named-block architectural protocol). *Let $\mathfrak{F}_{\mathrm{adm}}$ be an admissible representation class, and let $\mathcal{O}_{\mathrm{rs}}$ be an observation protocol on $\mathfrak{F}_{\mathrm{adm}}$. Let H be a finite-dimensional real vector space with a specified decomposition $H = H_0 \oplus H_1 \oplus \cdots \oplus H_B$. For $b = 0, \ldots, B$, let $P_b : H \to H$ be the canonical projection onto H_b along $\bigoplus_{c \neq b} H_c$. A named-block architectural protocol is a pair $(\mathcal{O}_{\mathrm{rs}}, \mathcal{T}_{\mathrm{rs}})$, where $\mathcal{T}_{\mathrm{rs}} = \big(C, U, (P_b)_{b=0}^B, (\ell_j)_{j=1}^m, (v_r)_{r=1}^n, ((H_b, \sigma_b))_{b \in B_{\mathrm{relu}}}\big)$ is a family of typed architectural data with the following components:*

(i) $C \in \mathrm{Sym}^2(H^*)$ is a positive definite symmetric bilinear form;

(ii) $U : H \to \mathbb{R}^k$ is a linear readout map;

(iii) $(P_b)_{b=0}^B$ is the named family of canonical projections;

(iv) $\ell_1, \ldots, \ell_m \in H^*$ are fixed linear measurements;

(v) $v_1, \ldots, v_n \in H$ are fixed state-space directions;

(vi) $B_{\mathrm{relu}} \subseteq \{1, \ldots, B\}$ indexes blocks with protocol-specified coordinatewise ReLU maps. For each $b \in B_{\mathrm{relu}}$, a coordinate isomorphism $H_b \simeq \mathbb{R}^{d_b}$ is fixed, and $\sigma_b : H_b \to H_b$ is the corresponding coordinatewise ReLU map on all of H_b.

Proposition 6 (Named-block stabilizer bound). *Let $(\mathcal{O}_{\mathrm{rs}}, \mathcal{T}_{\mathrm{rs}})$ be a named-block architectural protocol on $\mathfrak{F}_{\mathrm{adm}}$. Assume that the data in $\mathcal{T}_{\mathrm{rs}}$ are induced by a family of reparameterization-covariant architecture-defining data maps that are fixed on $\mathfrak{F}_{\mathrm{adm}}$ with the displayed values. Let*

$\mathcal{G}_{\mathrm{act}} = \mathcal{G}_{\mathrm{act}}(\mathcal{O}_{\mathrm{rs}}, \mathfrak{F}_{\mathrm{adm}})$. *Define* $\mathcal{G}_{\mathrm{rs}}$ *to be the set of all* $A \in \mathrm{GL}(H)$ *satisfying* $C(Ah, Ah') = C(h, h')$ *for every* $h, h' \in H, U \circ A = U, AP_b = P_b A$ *for* $b = 0, \ldots, B, A^* \ell_j = \ell_j$ *for* $j = 1, \ldots, m, Av_r = v_r$ *for* $r = 1, \ldots, n, A_b \in \mathsf{Mon}_{>0}(H_b)$ *for* $b \in B_{\mathrm{relu}}$, *where* $A_b = A|_{H_b}$ *and* $\mathsf{Mon}_{>0}(H_b)$ *is the positive monomial group in the fixed ReLU coordinates of* H_b. *Then* $\mathcal{G}_{\mathrm{rs}} = \mathcal{G}(\mathcal{T}_{\mathrm{rs}})$ *and* $\mathcal{G}_{\mathrm{act}} \leq \mathcal{G}_{\mathrm{rs}}$.

Moreover, $\mathcal{G}_{\mathrm{act}} = \mathcal{G}_{\mathrm{rs}}$ *if and only if both of the following conditions hold:*

(a) $\mathfrak{F}_{\mathrm{adm}}$ *is invariant under* $\mathcal{G}_{\mathrm{rs}}$*;*

(b) the stabilizer bound induced by $\mathcal{T}_{\mathrm{rs}}$ *is complete for* $\mathcal{O}_{\mathrm{rs}}$ *on* $\mathfrak{F}_{\mathrm{adm}}$ *in the sense of Definition 18.*

The exchangeable same-type block case is stated separately in Proposition 7.

Definition 21 (Linear representation)**.** *Let* G *be a group and* X *a vector space. A linear representation is a group homomorphism* $\rho : G \to \mathrm{GL}(X)$. *It is faithful if* $\ker(\rho) = \{e_G\}$, *where* e_G *is the identity of* G.

The named-block setting in Proposition 6 requires each block to be preserved individually, whereas the exchangeable-block setting considered below additionally allows permutations among blocks of the same type, followed by type-preserving transformations within each block. Therefore, both a type index and a block index are required.

Lemma 2 (Exchangeable-block semidirect-product representation)**.** *Let* H *be a finite-dimensional real vector space. Let* $\mathcal{T}_{\mathrm{blk}}$ *be a finite set of block types, and let* $B = \bigsqcup_{\tau \in \mathcal{T}_{\mathrm{blk}}} B_\tau$ *be a partition of the block index set by type. For every block* $b \in B$, *let* $H_b \leq H$ *be a linear subspace. Assume* $H = \bigoplus_{\tau \in \mathcal{T}_{\mathrm{blk}}} \bigoplus_{b \in B_\tau} H_b$.

For each type τ, *fix a common representation space* $V_\tau \neq \{0\}$ *for all blocks of type* τ, *together with fixed isomorphisms* $\iota_b : V_\tau \to H_b$ *for every block* $b \in B_\tau$. *Let* $G_\tau \leq \mathrm{GL}(V_\tau)$ *be the stabilizer of the architecture-defining block-local data fixed for block type* τ, *and let* $P_\tau \leq \mathrm{Sym}(B_\tau)$ *be the allowed permutation group for the blocks of type* τ.

Let $g = (g_b)_{b \in B_\tau} \in G_\tau^{B_\tau}$, *where each* g_b *specifies an allowed block-local transformation associated with each block* b. *Let* P_τ *act on* $G_\tau^{B_\tau}$ *by index permutation:* $(\pi \cdot g)_b = g_{\pi^{-1}(b)}$ *for every* $\pi \in P_\tau$ *and every* $b \in B_\tau$, *where* π *specifies an allowed permutation of the blocks of type* τ.

Define $\rho_\tau : G_\tau^{B_\tau} \rtimes P_\tau \longrightarrow \mathrm{GL}\left(\bigoplus_{b \in B_\tau} H_b\right)$. *For each* $(g, \pi) \in G_\tau^{B_\tau} \rtimes P_\tau$, *the operator* $\rho_\tau(g, \pi) : \bigoplus_{b \in B_\tau} H_b \longrightarrow \bigoplus_{b \in B_\tau} H_b$ *is defined to realize the combined action of the block-local transformations* g *and the block permutation* π *by* $[\rho_\tau(g, \pi)h]_b = \left(\iota_b \circ g_b \circ \iota_{\pi^{-1}(b)}^{-1}\right)\left(h_{\pi^{-1}(b)}\right)$ *for every* $g = (g_b)_{b \in B_\tau} \in G_\tau^{B_\tau}$, *every* $\pi \in P_\tau$, *every* $h = (h_b)_{b \in B_\tau} \in \bigoplus_{b \in B_\tau} H_b$, *and every* $b \in B_\tau$. *Here* $[\cdot]_b$ *denotes the component in* H_b. *Then* ρ_τ *is a faithful representation of* $G_\tau^{B_\tau} \rtimes P_\tau$ *on* $\bigoplus_{b \in B_\tau} H_b$.

Proposition 7 (Exchangeable-block semidirect-product bound)**.** *Adopt the setting and notation of Lemma 2. Let* $\mathcal{O}$ *be an observation protocol on* $\mathfrak{F}_{\mathrm{adm}}$, *and let* $\mathcal{G}_{\mathrm{act}} = \mathcal{G}_{\mathrm{act}}(\mathcal{O}, \mathfrak{F}_{\mathrm{adm}})$. *For each type* τ, *set* $\mathfrak{W}_\tau = \rho_\tau(G_\tau^{B_\tau} \rtimes P_\tau)$, *and let* $\mathfrak{W} = \prod_{\tau \in \mathcal{T}_{\mathrm{blk}}} \mathfrak{W}_\tau$ *act blockwise on* H. *Let* $\mathcal{T}_{\mathrm{glob}}$ *be a family of global architecture-defining data, not assigned to individual blocks, induced by reparameterization-covariant data maps that are fixed on* $\mathfrak{F}_{\mathrm{adm}}$, *and define* $\mathcal{G}_{\mathrm{glob}} = \mathcal{G}(\mathcal{T}_{\mathrm{glob}}), \mathcal{G}_{\mathrm{exch}} = \mathcal{G}_{\mathrm{glob}} \cap \mathfrak{W}$. *Assume that every* $A \in \mathcal{G}_{\mathrm{act}}$ *satisfies the following conditions for every* $\tau \in \mathcal{T}_{\mathrm{blk}}$*:*

(i) there exists $\pi_\tau \in P_\tau$ *such that* $A(H_b) = H_{\pi_\tau(b)}$ *for every* $b \in B_\tau$*;*

(ii) for every target block $b \in B_\tau$, $\iota_b^{-1} \circ A|_{H_{\pi_\tau^{-1}(b)}} \circ \iota_{\pi_\tau^{-1}(b)} \in G_\tau$.

Then $\mathcal{G}_{\mathrm{act}} \leq \mathcal{G}_{\mathrm{exch}}$. *Moreover,* $\mathcal{G}_{\mathrm{act}} = \mathcal{G}_{\mathrm{exch}}$ *if and only if* $\mathfrak{F}_{\mathrm{adm}}$ *is invariant under* $\mathcal{G}_{\mathrm{exch}}$ *and every* $A \in \mathcal{G}_{\mathrm{exch}}$ *satisfying* $A \cdot \mathfrak{F}_{\mathrm{adm}} = \mathfrak{F}_{\mathrm{adm}}$ *also satisfies* $\Phi_{\mathcal{O}}(A \cdot f) = \Phi_{\mathcal{O}}(f)$ *for every* $f \in \mathfrak{F}_{\mathrm{adm}}$.

If $P_\tau = \mathrm{Sym}(B_\tau)$, *then* $G_\tau^{B_\tau} \rtimes P_\tau \cong G_\tau \wr \mathrm{Sym}(B_\tau)$. *For a proper subgroup* $P_\tau < \mathrm{Sym}(B_\tau)$, *we retain the explicit semidirect-product notation.*

Example 3 (A blockwise obstruction for an actual symmetry group)**.** *Suppose a named-block protocol fixes the readout rows, a residual metric, and the head subspaces, but does not fix a basis inside*

a head subspace H_b. A one-dimensional subspace of H_b can be determined by the protocol only if it is fixed by the actual admissible observation-preserving group. In particular, if the action of that actual group on H_b contains the full C-orthogonal group and $\dim H_b \geq 2$, then no nonzero one-dimensional subspace of H_b is determined by the protocol. Containment of this orthogonal group in a stabilizer upper bound alone would not imply the conclusion.

B AN AUXILIARY ORBIT EXAMPLE

Example 4 (Orbit separation and fixed symbolic assignments)**.** *Let $\mathfrak{F} = \{a, b\}$, let $K = \{e\}$ act trivially, and take $\Phi_{\mathcal{O}}(a) = \Phi_{\mathcal{O}}(b)$. The observational fibre $\{a, b\}$ contains two distinct K-orbits, so orbit separation fails. Both assignments $\beta_0(a) = \beta_0(b) = 0$ and $\beta_1(a) = 0,\ \beta_1(b) = 1$ are K-invariant, but only β_0 is exactly extractable. Thus orbit separation is unnecessary for a particular invariant assignment, whereas invariance alone need not ensure extractability.*

C PROOFS FOR FINITE REPRESENTATION CALIBRATION

C.1 EXACT FIBRES AND THE FULL LINEAR SYMMETRY GROUP

Proof of Theorem 4. Write $N_{\mathrm{cal}} = L \cap S_{\mathrm{cal}}^{\perp}$. Then $L = S_{\mathrm{cal}} \oplus N_{\mathrm{cal}}$ is an orthogonal decomposition. An orthogonal transformation fixing every b_i fixes S_{cal} pointwise and preserves its orthogonal complement. Consequently every element of $\mathcal{K}_{\mathrm{cal}}$ has the unique form

$$A = I_{S_{\mathrm{cal}}} \oplus Q \oplus I_{L^{\perp}},\ Q \in O(N_{\mathrm{cal}}).$$

Conversely every such transformation is in $\mathcal{K}_{\mathrm{cal}}$. Restriction to N_{cal} is a group isomorphism, including when this subspace is zero-dimensional, when its orthogonal group is trivial.

For $h, h' \in L$, equality $B_{\mathrm{cal}}^{\top} h = B_{\mathrm{cal}}^{\top} h'$ is equivalent to $h - h' \perp S_{\mathrm{cal}}$, and hence to $P_{\mathrm{cal}} h = P_{\mathrm{cal}} h'$. Given this equality, $\|h\|_2 = \|h'\|_2$ is equivalent to equality of the norms of their components in N_{cal}. Two vectors of equal norm in a Euclidean space are related by an orthogonal transformation: for nonzero vectors, extend their normalized vectors to orthonormal bases and map one basis to the other; for two zero vectors use the identity. This also covers a one-dimensional complement. Extending that transformation as the identity on the other summands gives $h' = Ah$ with $A \in \mathcal{K}_{\mathrm{cal}}$. The reverse implication follows directly from the displayed decomposition. Thus the fibres are exactly the orbits.

Suppose now that $T \in \mathrm{GL}(L)$ preserves Φ_{cal} for every $h \in L$. Norm preservation and the real polarization identity imply that T is orthogonal on L. Moreover, $B_{\mathrm{cal}}^{\top} T h = B_{\mathrm{cal}}^{\top} h$ for every $h \in L$ implies $T^{\top} b_i = b_i$. Multiplying by T gives $T b_i = b_i$. Its identity extension to $L^{\perp}$ therefore lies in $\mathcal{K}_{\mathrm{cal}}$. This proves the full-group assertion without a separate completeness hypothesis.

If $u_w \in S_{\mathrm{cal}}$, write $u_w = B_{\mathrm{cal}} t$ for some coefficient vector t. For every $h \in L$, $w^{\top} h = t^{\top} B_{\mathrm{cal}}^{\top} h$, so its sign is determined by the observations. If $u_w \notin S_{\mathrm{cal}}$, then $v_{\mathrm{cal}} \neq 0$. The states v_{cal} and $-v_{\mathrm{cal}}$ have the same observed inner products, all zero, and the same norm, but their scores are respectively $\|v_{\mathrm{cal}}\|_2^2$ and $-\|v_{\mathrm{cal}}\|_2^2$. They witness non-identifiability.

Finally, write $h = P_{\mathrm{cal}} h + t_h$, where $t_h \in N_{\mathrm{cal}}$. Its fibre consists of $P_{\mathrm{cal}} h + t$ with $t \in N_{\mathrm{cal}}$ and $\|t\|_2 = \|t_h\|_2$. The score is $a_h + v_{\mathrm{cal}}^{\top} t$. Cauchy–Schwarz bounds the second term between $-R_h$ and R_h, with both extrema attained by taking t parallel or antiparallel to v_{cal} when $R_h > 0$. When $R_h = 0$, every score equals a_h. For $R_h > 0$, the extrema have a common nonzero sign exactly when $|a_h| > R_h$. If $|a_h| < R_h$ the extrema have opposite signs, and if $|a_h| = R_h$ one is zero and the other nonzero. These facts prove the fibrewise criterion. We use only the extrema, not an assertion that every intermediate score is attained, which would be false for a one-dimensional residual sphere. □

C.2 ATTAINED WORST-CASE DISAGREEMENT

Proof of Theorem 5. For two nonzero vectors $s, t \in L$ and $h \sim \mathcal{N}(0, P_L)$, rotational symmetry in their two-dimensional span gives

$$\Pr\{\mathrm{sign}(s^{\top} h) \neq \mathrm{sign}(t^{\top} h)\} = \frac{1}{\pi} \arccos \frac{s^{\top} t}{\|s\|_2 \|t\|_2}.$$

Indeed, the disagreement region consists of two planar sectors whose combined angle is twice the angle between the normals. The polar angle of a nondegenerate two-dimensional standard Gaussian is uniform. Collinear normals give probabilities zero or one directly, including when $r_L = 1$. All nonzero linear scores have zero probability of a tie. This is the standard random-hyperplane identity underlying sign embeddings (Plan & Vershynin, 2014).

Decompose $u_w = a_w + v_{\mathrm{cal}}$ with $a_w = P_{\mathrm{cal}} w$. For $A \in \mathcal{K}_{\mathrm{cal}}$, the decomposition in the preceding proof gives

$$A^\top u_w = a_w + Q^\top v_{\mathrm{cal}},\ u_w^\top A^\top u_w = \|a_w\|_2^2 + v_{\mathrm{cal}}^\top Q^\top v_{\mathrm{cal}}.$$

The latter expression is at least $\|a_w\|_2^2 - \|v_{\mathrm{cal}}\|_2^2$, by Cauchy–Schwarz. Orthogonality also gives $\|A^\top u_w\|_2 = \|u_w\|_2$. If $v_{\mathrm{cal}} \neq 0$, the reflection in Equation 5 fixes S_{cal} and $L^\perp$, reverses v_{cal}, and attains this lower bound. Thus the largest angle is

$$\arccos(1 - 2\eta_{\mathrm{cal}}^2) = 2 \arcsin \eta_{\mathrm{cal}},\ 0 \le \eta_{\mathrm{cal}} \le 1.$$

The Gaussian identity proves Equation 4. The reflection is symmetric and its square is the identity, so every singular value is one; its difference from the identity has rank one. If $v_{\mathrm{cal}} = 0$, every allowed A fixes u_w and hence all probe predictions on L. This proves every assertion. □

C.3 The finite-sample calibration law

Proof of Corollary 6. Identify L isometrically with $\mathbb{R}^{r_L}$. For $0 < n_{\mathrm{cal}} < r_L$, a matrix with independent standard Gaussian columns has rank n_{cal} almost surely. For example, the determinant of the leading square minor is a nonzero polynomial in continuously distributed entries, and hence vanishes with probability zero. Rotational invariance implies that its column span is a uniformly distributed n_{cal}-dimensional subspace. Because u_w was fixed independently of these columns, its normalized squared projection onto the orthogonal complement has the same law as

$$\frac{\sum_{j=1}^{r_L - n_{\mathrm{cal}}} g_j^2}{\sum_{j=1}^{r_L} g_j^2},\ (g_1, \ldots, g_{r_L}) \sim \mathcal{N}(0, I_{r_L}).$$

The numerator and the remaining sum are independent chi-squared variables with $r_L - n_{\mathrm{cal}}$ and n_{cal} degrees of freedom. Transforming their joint gamma density to the ratio and total yields the beta density with half these degrees of freedom as its parameters. This proves Equation 6. The function $x \mapsto (2/\pi) \arcsin \sqrt{x}$ is strictly increasing on $[0, 1]$, giving Equation 7.

If $n_{\mathrm{cal}} = 0$, then $S_{\mathrm{cal}} = \{0\}$ and $\eta_{\mathrm{cal}} = 1$. If $n_{\mathrm{cal}} \ge r_L$, the calibration states span L almost surely, so $\eta_{\mathrm{cal}} = 0$. In the intermediate case the beta distribution has no point mass at zero; therefore $u_w \notin S_{\mathrm{cal}}$ almost surely. Theorem 4 gives the exact-identification statement. □

C.4 Application and limitations

The deterministic results are conditional on the complete calibration matrix and do not require its columns to be independent or Gaussian. In contrast, the beta law does not apply when the probe is chosen after seeing those same columns. For example, choosing a nonzero probe inside their span makes $\eta_{\mathrm{cal}} = 0$ even with fewer columns than r_L. The Gaussian evaluation assumption is separate from the calibration-sampling assumption. The threshold $n_{\mathrm{cal}} \ge r_L$ is specific to independent Gaussian calibration, not a minimax lower bound for all observation designs. If arbitrary target-dependent calibration states can be selected, the single choice $b_1 = u_w$ already identifies this fixed probe.

For an attention head with column-vector convention, $W_V' = AW_V$, $b_V' = Ab_V$, and $W_O' = W_O A^{-1}$ leave its output unchanged whenever queries and keys are unchanged. Taking $A = A_{\mathrm{ref}}$ additionally fixes each selected last-token calibration context b_i; it does not assert that contexts at other tokens are fixed. Other architectural constraints can exclude this transformation. The example therefore applies only at sites where this compensated value-space action is admissible. It proves a negative witness whenever an actual evaluation probe changes, but does not characterize all models with the same language-model outputs.

Nor does a finite sample identify its population support. Any finite set lies in its span, even when the population has larger support. Replacing L by a principal subspace estimated from the calibration set would change both the admissible group and the population being studied. Such a replacement is not justified by the theorems.

C.5 THE SIGN-SYMMETRIC RANDOM-TRANSFORMATION BASELINE

Proposition 8 (Expected disagreement under sign-symmetric sampling). *Fix w, h with $w^\top h \neq 0$. Suppose a random matrix A has the same distribution as $-A$ and $\Pr\{w^\top Ah = 0\} = 0$. Then*

$$\Pr\{\operatorname{sign}(w^\top Ah) \neq \operatorname{sign}(w^\top h)\} = \frac{1}{2}.$$

The same expectation holds after averaging over any fixed finite set of nonzero baseline scores, without requiring independence between predictions on different states.

Proof. The transformed score has a distribution symmetric about zero. Since it has no atom at zero, its positive and negative probabilities are both $1/2$. Exactly one of these signs disagrees with the fixed original sign. The finite-set assertion follows from linearity of expectation. □

Haar-distributed orthogonal matrices satisfy the matrix sign symmetry. Products of independent Haar orthogonal factors with a fixed positive singular-value matrix do so as well. A numerical orthogonal-matrix generator must implement this distribution, including the appropriate QR sign correction, before this explanation is applied to it. The result is a conditional sampling baseline, not evidence of a loss of task information.

D EXACT OBSERVATION FACTORIZATION

Proof. Assume first that β is exactly extractable from $\mathcal{O}$. Let $E : \operatorname{im}(\Phi_\mathcal{O}) \to \mathfrak{S}/\!\sim_\mathfrak{S}$ be a correct observation-only extractor. If $f \equiv_\mathcal{O} g$, then $\Phi_\mathcal{O}(f) = \Phi_\mathcal{O}(g)$. Therefore $q_\mathfrak{S}(\beta(f)) = E(\Phi_\mathcal{O}(f)) = E(\Phi_\mathcal{O}(g)) = q_\mathfrak{S}(\beta(g))$. Hence $\beta(f) \sim_\mathfrak{S} \beta(g)$. This proves *(ii)*.

Assume *(ii)* and let $y \in \operatorname{im}(\Phi_\mathcal{O})$. If $f, g \in \Phi_\mathcal{O}^{-1}(\{y\})$, then $\Phi_\mathcal{O}(f) = \Phi_\mathcal{O}(g)$, so $f \equiv_\mathcal{O} g$. Condition *(ii)* gives $q_\mathfrak{S}(\beta(f)) = q_\mathfrak{S}(\beta(g))$. Thus $q_\mathfrak{S} \circ \beta$ is constant on $\Phi_\mathcal{O}^{-1}(\{y\})$. This proves *(iii)*.

Assume *(iii)*. For each $y \in \operatorname{im}(\Phi_\mathcal{O})$, define $E(y)$ to be the common value of $q_\mathfrak{S} \circ \beta$ on $\Phi_\mathcal{O}^{-1}(\{y\})$. The fibre is nonempty because $y \in \operatorname{im}(\Phi_\mathcal{O})$, and its common value is well defined by *(iii)*. For every $f \in \mathfrak{F}$, $E(\Phi_\mathcal{O}(f)) = q_\mathfrak{S}(\beta(f)) = [\beta(f)]_{\sim_\mathfrak{S}}$. Thus E is a correct observation-only extractor. This proves *(i)*.

Assume *(ii)*. Define a factor map $\overline{\beta}_\mathcal{O} : \mathfrak{F}/\!\equiv_\mathcal{O} \to \mathfrak{S}/\!\sim_\mathfrak{S}$ by $\overline{\beta}_\mathcal{O}([f]_{\equiv_\mathcal{O}}) = q_\mathfrak{S}(\beta(f))$. If $[f]_{\equiv_\mathcal{O}} = [g]_{\equiv_\mathcal{O}}$, then $f \equiv_\mathcal{O} g$. Condition *(ii)* therefore gives $q_\mathfrak{S}(\beta(f)) = q_\mathfrak{S}(\beta(g))$, so $\overline{\beta}_\mathcal{O}$ is well defined. $\forall f \in \mathfrak{F}$, $\overline{\beta}_\mathcal{O}(\pi_\mathcal{O}(f)) = q_\mathfrak{S}(\beta(f))$. Hence $\overline{\beta}_\mathcal{O} \circ \pi_\mathcal{O} = q_\mathfrak{S} \circ \beta$. Since $\pi_\mathcal{O}$ is surjective, this factor map is unique. This proves *(iv)*.

Conversely, assume *(iv)*. If $f \equiv_\mathcal{O} g$, then $\pi_\mathcal{O}(f) = \pi_\mathcal{O}(g)$. Therefore $q_\mathfrak{S}(\beta(f)) = \overline{\beta}_\mathcal{O}(\pi_\mathcal{O}(f)) = \overline{\beta}_\mathcal{O}(\pi_\mathcal{O}(g)) = q_\mathfrak{S}(\beta(g))$. Hence $\beta(f) \sim_\mathfrak{S} \beta(g)$. This proves *(ii)*.

To prove uniqueness of the extractor, let E' be another correct observation-only extractor. For any $y \in \operatorname{im}(\Phi_\mathcal{O})$, choose $f \in \mathfrak{F}$ with $\Phi_\mathcal{O}(f) = y$. Then $E'(y) = q_\mathfrak{S}(\beta(f)) = E(y)$. Thus $E' = E$.

We say the map $\widetilde{\Phi}_\mathcal{O}$ is well defined because $f \equiv_\mathcal{O} g$ implies $\Phi_\mathcal{O}(f) = \Phi_\mathcal{O}(g)$. It is injective because equality of observations implies observational equivalence, and it is surjective by the definition of $\operatorname{im}(\Phi_\mathcal{O})$. Finally, $\overline{\beta}_\mathcal{O} = E \circ \widetilde{\Phi}_\mathcal{O}$, which is equivalent to $E = \overline{\beta}_\mathcal{O} \circ \widetilde{\Phi}_\mathcal{O}^{-1}$. □

E ADDITIONAL PROOFS

E.1 MONOTONICITY UNDER OBSERVATION REFINEMENT

Proof. If $f \equiv_{\mathcal{O}_2} g$, then $\Phi_{\mathcal{O}_1}(f) = r_{21}(\Phi_{\mathcal{O}_2}(f)) = r_{21}(\Phi_{\mathcal{O}_2}(g)) = \Phi_{\mathcal{O}_1}(g)$, which proves *(i)*. Part *(i)* also shows that κ_{21} is well defined. It is surjective because every class $[f]_{\equiv_{\mathcal{O}_1}}$ is the image of $[f]_{\equiv_{\mathcal{O}_2}}$.

If E_1 is correct for β, then for every $f \in \mathfrak{F}$, $(E_1 \circ r_{21})(\Phi_{\mathcal{O}_2}(f)) = E_1(\Phi_{\mathcal{O}_1}(f)) = [\beta(f)]_{\sim_\mathfrak{S}}$. This proves *(iii)*. □

E.2 OBJECTWISE SYMMETRY OBSTRUCTION

Proof. Set $h = g \cdot f$. Since $g \in \Gamma_{\mathcal{O}}(f)$, Definition 6 gives $\Phi_{\mathcal{O}}(h) = \Phi_{\mathcal{O}}(g \cdot f) = \Phi_{\mathcal{O}}(f)$. Hence $h \equiv_{\mathcal{O}} f$. By assumption, $\beta(h) = \beta(g \cdot f) \not\sim_{\mathfrak{S}} \beta(f)$. Therefore (h, f) is a non-identifiability witness. Corollary 1 gives the conclusion. □

E.3 BEHAVIORAL REPARAMETERIZATION WITNESS

Proof. Write $\theta = (h, W)$. For every $x \in \mathcal{X}$, $y_{A\cdot\theta}(x) = (W \circ A^{-1})(Ah(x)) = Wh(x) = y_\theta(x)$. Therefore $\Phi_{\mathrm{beh}}(A \cdot \theta) = \Phi_{\mathrm{beh}}(\theta)$, and hence $A \cdot \theta \equiv_{\mathcal{O}_{\mathrm{beh}}} \theta$.

Now suppose $\beta(A \cdot \theta) \not\sim_{\mathfrak{S}} \beta(\theta)$. Then $(A \cdot \theta, \theta)$ is a non-identifiability witness. By Corollary 1, β is not exactly extractable from $\mathcal{O}_{\mathrm{beh}}$. □

E.4 WITNESS CHARACTERIZATION OF NON-IDENTIFIABILITY

Proof. Let (f, g) be a non-identifiability witness. Then $f \equiv_{\mathcal{O}} g$ and $\beta(f) \not\sim_{\mathfrak{S}} \beta(g)$. Hence condition *(ii)* of Proposition 1 fails, so β is not exactly extractable from $\mathcal{O}$.

Conversely, suppose that β is not exactly extractable from $\mathcal{O}$. By Proposition 1, condition *(ii)* fails. Therefore, there exist $f, g \in \mathfrak{F}$ such that $f \equiv_{\mathcal{O}} g$ and $\beta(f) \not\sim_{\mathfrak{S}} \beta(g)$. Hence (f, g) is a non-identifiability witness for β with respect to $\mathcal{O}$. □

E.5 MONOTONICITY OF GLOBAL OBSERVATION-PRESERVING SYMMETRIES

Proof. Let $g \in \mathcal{G}_{\mathcal{O}_2}$. For every $f \in \mathfrak{F}$, $\Phi_{\mathcal{O}_1}(g \cdot f) = r_{21}(\Phi_{\mathcal{O}_2}(g \cdot f)) = r_{21}(\Phi_{\mathcal{O}_2}(f)) = \Phi_{\mathcal{O}_1}(f)$. Hence $g \in \mathcal{G}_{\mathcal{O}_1}$. □

E.6 OBSERVATION-PRESERVING SYMMETRIES FORM A SUBGROUP

Proof. The identity belongs to $\mathcal{G}_{\mathcal{O}}$ because $\Phi_{\mathcal{O}}(e \cdot f) = \Phi_{\mathcal{O}}(f)$ for every $f \in \mathfrak{F}$.

Let $g, h \in \mathcal{G}_{\mathcal{O}}$. For every $f \in \mathfrak{F}$, we have $h \cdot f \in \mathfrak{F}$, and hence $\Phi_{\mathcal{O}}((gh) \cdot f) = \Phi_{\mathcal{O}}(g \cdot (h \cdot f)) = \Phi_{\mathcal{O}}(h \cdot f) = \Phi_{\mathcal{O}}(f)$. Therefore $gh \in \mathcal{G}_{\mathcal{O}}$.

Finally, let $g \in \mathcal{G}_{\mathcal{O}}$ and $f \in \mathfrak{F}$. Set $u = g^{-1} \cdot f$. Then $u \in \mathfrak{F}$ and $f = g \cdot u$. Therefore $\Phi_{\mathcal{O}}(f) = \Phi_{\mathcal{O}}(g \cdot u) = \Phi_{\mathcal{O}}(u) = \Phi_{\mathcal{O}}(g^{-1} \cdot f)$. Hence $g^{-1} \in \mathcal{G}_{\mathcal{O}}$. Thus $\mathcal{G}_{\mathcal{O}} \leq \mathcal{G}$. □

E.7 ORBIT SPECIALIZATION OF EXACT EXTRACTABILITY

Proof. By Proposition 1, β is exactly extractable if and only if it is constant modulo $\sim_{\mathfrak{S}}$ on every observational equivalence class. Orbit separation relative to K states that these equivalence classes are exactly the K-orbits. Therefore fibre constancy is equivalent to $\beta(k \cdot f) \sim_{\mathfrak{S}} \beta(f)$ for every $k \in K$ and every $f \in \mathfrak{F}$. □

E.8 UNIVERSAL CHARACTERIZATION OF ORBIT SEPARATION

Proof. Assume first that $\mathcal{O}$ is orbit-separating relative to K. For every symbolic assignment $\beta : \mathfrak{F} \to \{0, 1\}$, Corollary 4 gives β is exactly extractable from $\mathcal{O} \iff (\beta(k \cdot f) = \beta(f)$ for every $k \in K$ and every $f \in \mathfrak{F})$. Hence *(i)* implies *(ii)*.

Conversely, assume *(ii)*. Since $K \leq \mathcal{G}_{\mathcal{O}}$, one has $h = k \cdot f \Longrightarrow h \equiv_{\mathcal{O}} f$ for every $k \in K$ and every $f \in \mathfrak{F}$. Thus every K-orbit is contained in an observational equivalence class.

Suppose, for contradiction, that $\mathcal{O}$ is not orbit-separating relative to K. Then there exist $f, h \in \mathfrak{F}$ such that $f \equiv_{\mathcal{O}} h$ but $h \notin K \cdot f$. Define $\beta : \mathfrak{F} \to \{0, 1\}$ by $\beta(x) = \begin{cases} 1, & x \in K \cdot f, \\ 0, & x \notin K \cdot f. \end{cases}$ Both $K \cdot f$ and its complement are invariant under the action of K, so $\beta(k \cdot x) = \beta(x)$ for every $k \in K$ and every $x \in \mathfrak{F}$. Condition *(ii)* therefore implies that β is exactly extractable from $\mathcal{O}$. However, $f \equiv_{\mathcal{O}} h$ while $\beta(f) = 1$ and $\beta(h) = 0$, contradicting Proposition 1. Therefore $f \equiv_{\mathcal{O}} h \Longrightarrow h \in K \cdot f$. The

reverse implication follows from $K \leq \mathcal{G}_{\mathcal{O}}$. Hence the observational equivalence classes are exactly the K-orbits, and $\mathcal{O}$ is orbit-separating relative to K. Thus *(ii)* implies *(i)*. □

E.9 Fixed homogeneous sign-probe gauge criterion

Proof. By Proposition 1, β_w is exactly extractable from $\mathcal{O}$ iff $\beta_w(h) = \beta_w(h')$ whenever $h \equiv_{\mathcal{O}} h'$. By the assumed characterization of observational equivalence, this is equivalent to $\operatorname{sign}(w^\top Ah) = \operatorname{sign}(w^\top h)$ for every $A \in \mathcal{G}$ and every $h \in \mathcal{H}$. The three sign fibres partition $\mathcal{H}$, so we have the orbit-containment statement.

Assume $\mathcal{H} = \mathbb{R}^d$ and fix $A \in \mathcal{G}$. Define the nonzero linear functionals $q(h) = w^\top h$ and $q_A(h) = w^\top Ah = (A^\top w)^\top h$. If $\operatorname{sign}(q_A(h)) = \operatorname{sign}(q(h))$ for every $h \in \mathbb{R}^d$, then $q(h) = 0$ iff $q_A(h) = 0$. Hence $\ker q = \ker q_A$. Two nonzero linear functionals on a finite-dimensional vector space with the same kernel are scalar multiples, so there exists a unique $\chi(A) \in \mathbb{R} \setminus \{0\}$ such that $q_A = \chi(A)q$. Equivalently, $A^\top w = \chi(A)w$.

Choose $h_+ \in \mathbb{R}^d$ such that $q(h_+) > 0$. Sign preservation gives $q_A(h_+) > 0$, and therefore $\chi(A) > 0$.

Conversely, if $A^\top w = \chi(A)w$ for some $\chi(A) > 0$, then $\operatorname{sign}(w^\top Ah) = \operatorname{sign}(\chi(A)w^\top h) = \operatorname{sign}(w^\top h)$ for every $h \in \mathbb{R}^d$. □

E.10 Full-$\mathrm{GL}(d)$ fixed homogeneous sign-probe obstruction

Proof. Since $w \neq 0$, $\exists h_+, h_- \in \mathbb{R}^d$ such that $w^\top h_+ > 0$ and $w^\top h_- < 0$. Hence $\beta_w(h_+) = 1$ and $\beta_w(h_-) = -1$. Because $\mathrm{GL}(d)$ acts transitively on nonzero vectors, there exists $A \in \mathrm{GL}(d)$ such that $Ah_+ = h_-$. Thus h_+ and h_- lie in the same $\mathrm{GL}(d)$-orbit. By the assumed orbit characterization of observational equivalence, this gives $h_+ \equiv_{\mathcal{O}} h_-$. However, $\beta_w(h_+) = 1 \neq -1 = \beta_w(h_-)$. Therefore (h_+, h_-) is a non-identifiability witness for β_w with respect to $\mathcal{O}$. Thus β_w is not exactly extractable from $\mathcal{O}$ by Corollary 1. □

E.11 Boundary-local fixed homogeneous sign-probe gauge criterion

Proof. Fix $A \in \mathrm{GL}(d)$ and assume *(i)*. Define the nonzero linear functionals $q(h) = w^\top h$ and $r(h) = w^\top Ah = (A^\top w)^\top h$. Because $q(h_0) = 0$ and $\operatorname{sign}(r(h_0)) = \operatorname{sign}(q(h_0)) = 0$, we have $r(h_0) = 0$.

Let $v \in \ker q$. Choose a nonzero scalar t sufficiently small that $h_0 + tv \in B^\circ_\varepsilon(h_0) \subseteq \mathcal{H}$. Because $v \in \ker q$ and $q(h_0) = 0$, we have $q(h_0 + tv) = 0$. The assumed sign equality therefore gives $r(h_0 + tv) = 0$. Using $r(h_0) = 0$, we obtain $0 = r(h_0 + tv) = r(h_0) + tr(v) = tr(v)$. Since $t \neq 0$, it follows that $r(v) = 0$. Thus $\ker q \subseteq \ker r$.

Because $w \neq 0$ and $A \in \mathrm{GL}(d)$, both q and r are nonzero linear functionals. Their kernels therefore both have codimension one. Hence $\ker q = \ker r$. Two nonzero linear functionals with the same kernel are scalar multiples, so there exists $\chi_A \in \mathbb{R} \setminus \{0\}$ such that $r = \chi_A q$. Equivalently, $A^\top w = \chi_A w$. Choose $u \in \mathbb{R}^d$ such that $q(u) > 0$. Choose $t > 0$ sufficiently small that $h_0 + tu \in B^\circ_\varepsilon(h_0)$. Then $q(h_0 + tu) = tq(u) > 0$. By the assumed sign equality, $r(h_0 + tu) > 0$. But $r(h_0 + tu) = \chi_A q(h_0 + tu)$. Because $q(h_0 + tu) > 0$, we have $\chi_A > 0$. Since $w \neq 0$, the scalar χ_A is unique. Thus *(i)* proves *(ii)*.

Conversely, assume *(ii)*. For every $h \in \mathcal{H}$, $w^\top Ah = (A^\top w)^\top h = \chi_A w^\top h$. Because $\chi_A > 0$, $\operatorname{sign}(w^\top Ah) = \operatorname{sign}(w^\top h)$. Thus *(ii)* proves *(i)*.

Finally, apply the equivalence separately to every $A \in \mathcal{G}$. The uniqueness of each scalar χ_A defines a unique map $\chi : \mathcal{G} \to \mathbb{R}_{>0}$ such that $A^\top w = \chi(A)w$ for every $A \in \mathcal{G}$. Conversely, any such map gives *(ii)* for every $A \in \mathcal{G}$, and hence *(i)* for every $A \in \mathcal{G}$. □

E.12 Fixed-data architectural stabilizer bound

Proof. We first verify that $\mathcal{G}_{\mathrm{act}} \leq \mathrm{GL}(H)$. The identity preserves both $\mathfrak{F}_{\mathrm{adm}}$ and $\Phi_{\mathcal{O}}$. If $A, B \in \mathcal{G}_{\mathrm{act}}$, then $(AB) \cdot \mathfrak{F}_{\mathrm{adm}} = A \cdot (B \cdot \mathfrak{F}_{\mathrm{adm}}) = \mathfrak{F}_{\mathrm{adm}}$, and, for every $f \in \mathfrak{F}_{\mathrm{adm}}$, $\Phi_{\mathcal{O}}((AB) \cdot f) = \Phi_{\mathcal{O}}(A \cdot (B \cdot f)) = \Phi_{\mathcal{O}}(B \cdot f) = \Phi_{\mathcal{O}}(f)$. If $A \in \mathcal{G}_{\mathrm{act}}$, then $A^{-1} \cdot \mathfrak{F}_{\mathrm{adm}} = \mathfrak{F}_{\mathrm{adm}}$. For $f \in \mathfrak{F}_{\mathrm{adm}}$, set

$g = A^{-1} \cdot f$. Since $g \in \mathfrak{F}_{\mathrm{adm}}$ and $A \cdot g = f$, $\Phi_{\mathcal{O}}(A^{-1} \cdot f) = \Phi_{\mathcal{O}}(g) = \Phi_{\mathcal{O}}(A \cdot g) = \Phi_{\mathcal{O}}(f)$. Thus $\mathcal{G}_{\mathrm{act}}$ is a subgroup of $\mathrm{GL}(H)$.

We now prove *(i)*. Let $A \in \mathcal{G}_{\mathrm{act}}$ and fix $j \in J$. Choose $f \in \mathfrak{F}_{\mathrm{adm}}$. Since A preserves the admissible class, $A \cdot f \in \mathfrak{F}_{\mathrm{adm}}$. The fixed-data condition and reparameterization covariance give $A \cdot T_j = A \cdot \tau_j(f) = \tau_j(A \cdot f) = T_j$. Hence $A \in \mathrm{Stab}_{\mathrm{GL}(H)}(T_j)$ for every $j \in J$. Therefore $A \in \mathcal{G}(\mathcal{T})$, and consequently $\mathcal{G}_{\mathrm{act}} \leq \mathcal{G}(\mathcal{T})$. The inclusion $\mathcal{G}(\mathcal{T}) \leq \mathrm{GL}(H)$ follows from Definition 12.

For *(ii)*, first assume $\mathcal{G}_{\mathrm{act}} = \mathcal{G}(\mathcal{T})$. Every element of $\mathcal{G}_{\mathrm{act}}$ maps $\mathfrak{F}_{\mathrm{adm}}$ onto itself, so $\mathfrak{F}_{\mathrm{adm}}$ is invariant under $\mathcal{G}(\mathcal{T})$. Moreover, every element of $\mathcal{G}(\mathcal{T})$ belongs to $\mathcal{G}_{\mathrm{act}}$ and therefore preserves the complete observation map on $\mathfrak{F}_{\mathrm{adm}}$. Thus the stabilizer bound is complete.

Conversely, assume invariance and completeness. Let $A \in \mathcal{G}(\mathcal{T})$. Invariance gives $A \cdot \mathfrak{F}_{\mathrm{adm}} = \mathfrak{F}_{\mathrm{adm}}$, and completeness gives $\Phi_{\mathcal{O}}(A \cdot f) = \Phi_{\mathcal{O}}(f)$ for every $f \in \mathfrak{F}_{\mathrm{adm}}$. Hence $A \in \mathcal{G}_{\mathrm{act}}$. Therefore $\mathcal{G}(\mathcal{T}) \leq \mathcal{G}_{\mathrm{act}}$. Combining this inclusion with *(i)* proves equality.

□

E.13 Stabilizer bound for model-dependent architectural data

Proof. Let $A \in \mathcal{G}_{\mathrm{act}}$, let $f \in \mathfrak{F}_{\mathrm{adm}}$, and fix $j \in J$. Reparameterization covariance and observation of τ_j give $A \cdot \tau_j(f) = \tau_j(A \cdot f) = \widehat{\tau}_j(\Phi_{\mathcal{O}}(A \cdot f)) = \widehat{\tau}_j(\Phi_{\mathcal{O}}(f)) = \tau_j(f)$. Hence A fixes every value $\tau_j(f)$, proving the upper bound.

If equality holds, invariance and completeness follow from the definition of $\mathcal{G}_{\mathrm{act}}$. Conversely, invariance and completeness imply that every element of the joint stabilizer of attained data values belongs to $\mathcal{G}_{\mathrm{act}}$. The upper bound then gives equality. □

E.14 Metric and readout stabilizer bound

Proof. By Definition 11, the metric is fixed exactly when $A^\top C A = C$, and the readout is fixed exactly when $U \circ A = U$. The latter condition is equivalent to $A^\top u_j = u_j$ for all readout covectors u_j, $j = 1, \ldots, m$. Hence $\mathcal{G}(C, U) = \{A \in \mathrm{GL}(d) : A^\top C A = C,\ A^\top u_j = u_j \text{ for } j = 1, \ldots, m\}$, and Theorem 6(i) gives $\mathcal{G}_{\mathrm{act}} \leq \mathcal{G}(C, U)$. With only the metric fixed, the stabilizer is $O(C) = \{A \in \mathrm{GL}(d) : A^\top C A = C\}$, which equals $O(d)$ when $C = I$. □

E.15 Global linear commutant of coordinatewise ReLU

Proof. Write $A = (a_{ji})_{j,i=1}^{d}$, and write the jth row of A as $a_j^\top$. The equation $\sigma(Az) = A\sigma(z)$ holds for every $z \in \mathbb{R}^d$ iff for every $j = 1, \ldots, d$ and every $z \in \mathbb{R}^d$, $\max\{a_j^\top z, 0\} = \sum_{i=1}^{d} a_{ji} \max\{z_i, 0\}$.

First, fix $i, j \in \{1, \ldots, d\}$ and take $z = -e_i$. Then $\sigma(z) = 0$, and the preceding equation gives $\max\{-a_{ji}, 0\} = 0$. Thus $a_{ji} \geq 0$. Since i, j were arbitrary, $a_{ji} \geq 0 (1 \leq i, j \leq d)$.

We next show that each row of A has at most one nonzero entry. For contradiction, assume that the jth row has two positive entries. Then there exist distinct indices $i \neq k$ such that $a_{ji} > 0$ and $a_{jk} > 0$. Define $z \in \mathbb{R}^d$ by $z_i = a_{jk}, z_k = -a_{ji}, z_\ell = 0 (\ell \notin \{i, k\})$. Then $a_j^\top z = a_{ji} a_{jk} - a_{jk} a_{ji} = 0$. Hence the left-hand side of the coordinate identity is $\max\{a_j^\top z, 0\} = 0$. On the other hand, $\sum_{\ell=1}^{d} a_{j\ell} \max\{z_\ell, 0\} = a_{ji} a_{jk} > 0$, contradicting the coordinate identity. Therefore each row of A has at most one nonzero entry.

Because $A \in \mathrm{GL}(d)$, no row of A is zero. Hence each row has exactly one nonzero entry. Since there are d rows, A has exactly d nonzero entries in total. Since A is invertible, no column of A is zero. Therefore each column also has exactly one nonzero entry. All nonzero entries are positive, so their positions determine a permutation $\pi \in S_d$, while their values determine a positive diagonal matrix $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. Hence $A = D\Pi_\pi$ with $\lambda_i > 0$ for every $i = 1, \ldots, d$, and therefore $A \in \mathsf{Mon}_{>0}(d)$.

Conversely, let $A = D\Pi_\pi \in \mathsf{Mon}_{>0}(d), D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \lambda_i > 0$. For every $x \in \mathbb{R}^d$, $\sigma(Dx) = D\sigma(x)$, because each λ_i is positive. Also, $\sigma(\Pi_\pi x) = \Pi_\pi \sigma(x)$, because Π_π only permutes

coordinates. Therefore, for every $z \in \mathbb{R}^d$, $\sigma(Az) = \sigma(D\Pi_\pi z) = D\sigma(\Pi_\pi z) = D\Pi_\pi \sigma(z) = A\sigma(z)$. Hence $\mathsf{Mon}_{>0}(d) \subseteq \{A \in \mathrm{GL}(d) : \sigma(Az) = A\sigma(z) \text{ for every } z \in \mathbb{R}^d\}$. Combining the two inclusions gives the desired equality. □

E.16 Fixed homogeneous sign-probe targets under reduced gauges

Proof. Assume that β_w is exactly extractable. Observation preservation and Proposition 1 give $\beta_w(Ah) = \beta_w(h)$ for every $A \in \mathcal{G}_{\mathrm{act}}$ and $h \in \mathcal{H}$. Theorem 3 applies under the stated boundary-local assumption and also when $\mathcal{H} = H$, by taking $h_0 = 0$. Thus each A has a unique $\chi_w(A) > 0$ satisfying $A^\top w = \chi_w(A)w$. For $A, B \in \mathcal{G}_{\mathrm{act}}$, $(AB)^\top w = B^\top A^\top w = \chi_w(A)\chi_w(B)w$. Uniqueness gives $\chi_w(AB) = \chi_w(A)\chi_w(B)$ and $\chi_w(I) = 1$, so χ_w is a group homomorphism.

For *(i)*, take $A = (-I_V) \oplus I_{V_c} \in \mathcal{G}_{\mathrm{act}}$. Since w vanishes on V_c, $A^\top w = -w$, contradicting the positive-ray condition for the nonzero covector w. For *(ii)*, the same argument uses $A = (-I_{H_b}) \oplus I_{H_{\neg b}} \in \mathcal{G}_{\mathrm{act}}$ and $w|_{H_{\neg b}} = 0$.

For *(iii)*, commutation with P_b makes both H_b and $H_{\neg b}$ invariant. Write $A|_{H_b} = D\Pi_\pi$, where $D = \mathrm{diag}(\lambda_1, \dots, \lambda_{d_b})$ with positive entries. The convention $\Pi_\pi e_j = e_{\pi(j)}$ gives $(D\Pi_\pi)^\top e_i^* = \lambda_i e^*_{\pi^{-1}(i)}$. Since e_i^* vanishes on $H_{\neg b}$, the same identity holds for $A^\top e_i^*$ after zero extension to H. Every $\pi \in \mathfrak{P}_b$ is realized by such an A. Exact extractability therefore requires $\pi(i) = i$ for every $\pi \in \mathfrak{P}_b$.

Conversely, if every $\pi \in \mathfrak{P}_b$ fixes i, then $A^\top e_i^* = \lambda_i e_i^*$ with $\lambda_i > 0$ for every $A \in \mathcal{G}_{\mathrm{act}}$. Hence $\beta_{e_i^*}$ is invariant under $\mathcal{G}_{\mathrm{act}}$. Under the additional orbit-separation assumption, Corollary 4 gives exact extractability. □

E.17 Named-block stabilizer bound

Proof. By assumption, $\mathcal{T}_{\mathrm{rs}}$ is fixed on $\mathfrak{F}_{\mathrm{adm}}$ through the stated family of reparameterization-covariant architecture-defining data maps. We first identify its joint stabilizer.

Let $A \in \mathrm{GL}(H)$. By the natural actions in Definition 11, the bilinear-form datum C is stabilized by A if and only if $C(Ah, Ah') = C(h, h')$ for every $h, h' \in H$. The readout datum U is stabilized if and only if $U \circ A = U$. The named projection family is stabilized if and only if $AP_b = P_b A$ for every $b = 0, \dots, B$. Since A is invertible, these projection equations imply $A(H_b) = H_b$ for every block H_b, so each restriction $A_b = A|_{H_b}$ belongs to $\mathrm{GL}(H_b)$.

For each fixed covector ℓ_j, the stabilizer equation is $A^*\ell_j = \ell_j$, and for each fixed vector v_r, it is $Av_r = v_r$. Now fix $b \in B_{\mathrm{relu}}$. Because $A(H_b) = H_b$, the nonlinear datum (H_b, σ_b) is stabilized if and only if $\sigma_b \circ A_b = A_b \circ \sigma_b$ on all of H_b. Theorem 7 gives $\sigma_b \circ A_b = A_b \circ \sigma_b \iff A_b \in \mathsf{Mon}_{>0}(H_b)$. Collecting these stabilizer equations yields $\mathcal{G}(\mathcal{T}_{\mathrm{rs}}) = \mathcal{G}_{\mathrm{rs}}$.

Since $\mathcal{T}_{\mathrm{rs}}$ is fixed on the admissible class through reparameterization-covariant data maps, Theorem 6(i) gives $\mathcal{G}_{\mathrm{act}} \leq \mathcal{G}(\mathcal{T}_{\mathrm{rs}}) = \mathcal{G}_{\mathrm{rs}}$.

It remains to prove the equality criterion. Suppose first that $\mathcal{G}_{\mathrm{act}} = \mathcal{G}_{\mathrm{rs}}$. Every element of $\mathcal{G}_{\mathrm{act}}$ maps $\mathfrak{F}_{\mathrm{adm}}$ onto itself, so $\mathfrak{F}_{\mathrm{adm}}$ is invariant under $\mathcal{G}_{\mathrm{rs}}$. Moreover, every element of $\mathcal{G}_{\mathrm{rs}} = \mathcal{G}_{\mathrm{act}}$ preserves the complete observation map on $\mathfrak{F}_{\mathrm{adm}}$. Hence the stabilizer bound induced by $\mathcal{T}_{\mathrm{rs}}$ is complete for $\mathcal{O}_{\mathrm{rs}}$ on $\mathfrak{F}_{\mathrm{adm}}$.

Conversely, assume that $\mathfrak{F}_{\mathrm{adm}}$ is invariant under $\mathcal{G}_{\mathrm{rs}}$ and that the stabilizer bound induced by $\mathcal{T}_{\mathrm{rs}}$ is complete for $\mathcal{O}_{\mathrm{rs}}$ on $\mathfrak{F}_{\mathrm{adm}}$. Let $A \in \mathcal{G}_{\mathrm{rs}}$. Invariance gives $A \cdot \mathfrak{F}_{\mathrm{adm}} = \mathfrak{F}_{\mathrm{adm}}$, and completeness gives $\Phi_{\mathcal{O}_{\mathrm{rs}}}(A \cdot f) = \Phi_{\mathcal{O}_{\mathrm{rs}}}(f)$ for every $f \in \mathfrak{F}_{\mathrm{adm}}$. Therefore $A \in \mathcal{G}_{\mathrm{act}}$, so $\mathcal{G}_{\mathrm{rs}} \leq \mathcal{G}_{\mathrm{act}}$. Combining this inclusion with the previously proved upper bound gives $\mathcal{G}_{\mathrm{act}} = \mathcal{G}_{\mathrm{rs}}$. This proves both directions of the equivalence. □

E.18 Exchangeable-block semidirect-product representation

Proof. Fix a type τ and define the linear isomorphism $J_\tau : \bigoplus_{b \in B_\tau} H_b \to V_\tau^{B_\tau}$ by $[J_\tau h]_b = \iota_b^{-1}(h_b)$, with inverse $[J_\tau^{-1}x]_b = \iota_b(x_b)$. Set $\widetilde{\rho}_\tau(g, \pi) = J_\tau \rho_\tau(g, \pi) J_\tau^{-1}$. The defining action gives $[\widetilde{\rho}_\tau(g, \pi)x]_b = g_b x_{\pi^{-1}(b)}$, so each operator is linear.

The semidirect-product multiplication is $(g,\pi)(g',\varpi) = (g(\pi \cdot g'), \pi\varpi)$, where $(\pi \cdot g')_b = g'_{\pi^{-1}(b)}$ (Dummit & Foote, 2004). For every $x \in V_\tau^{B_\tau}$ and $b \in B_\tau$, the coordinate formula yields $[\widetilde{\rho}_\tau(g,\pi)\widetilde{\rho}_\tau(g',\varpi)x]_b = g_b g'_{\pi^{-1}(b)} x_{\varpi^{-1}\pi^{-1}(b)} = [\widetilde{\rho}_\tau(g(\pi \cdot g'), \pi\varpi)x]_b$. Thus $\widetilde{\rho}_\tau$ preserves multiplication. It maps the identity to the identity operator, and $(g,\pi)^{-1} = (\pi^{-1} \cdot g^{-1}, \pi^{-1})$, with $g^{-1} = (g_b^{-1})_{b\in B_\tau}$, supplies an inverse for every represented operator. Since $\rho_\tau(g,\pi) = J_\tau^{-1}\widetilde{\rho}_\tau(g,\pi)J_\tau$, ρ_τ is also a linear representation.

For faithfulness, suppose $\rho_\tau(g,\pi) = I$, so $\widetilde{\rho}_\tau(g,\pi) = I$. If $B_\tau = \varnothing$, the domain group is trivial. Otherwise, fix $c \in B_\tau$ and choose a vector x whose only nonzero component is $x_c \in V_\tau$, using $V_\tau \neq \{0\}$. Its image has its only nonzero component at $\pi(c)$ because $g_{\pi(c)}$ is invertible. Equality of the image with x forces $\pi(c) = c$. Since c was arbitrary, $\pi = \mathrm{id}_{B_\tau}$. The coordinate formula now gives $g_b x_b = x_b$ for every $x_b \in V_\tau$, hence $g_b = I_{V_\tau}$ for every b. The kernel is therefore trivial, proving faithfulness. □

E.19 Exchangeable-block semidirect-product bound

Proof. Let $A \in \mathcal{G}_{\mathrm{act}}$. By Theorem 6(i), applied to the fixed global data $\mathcal{T}_{\mathrm{glob}}$, one has $A \in \mathcal{G}_{\mathrm{glob}}$. Fix a block type τ. By assumption, there exists $\pi_\tau \in P_\tau$ such that $A(H_b) = H_{\pi_\tau(b)}$ for every $b \in B_\tau$. For every target block $b \in B_\tau$, define $g_b = \iota_b^{-1} \circ A|_{H_{\pi_\tau^{-1}(b)}} \circ \iota_{\pi_\tau^{-1}(b)}$. The second assumption gives $g_b \in G_\tau$. Let $A_\tau = A|_{\bigoplus_{b\in B_\tau} H_b}$.

Under the coordinate identification $J_\tau : \bigoplus_{b\in B_\tau} H_b \longrightarrow V_\tau^{B_\tau}$, for every $x = (x_b)_{b\in B_\tau} \in V_\tau^{B_\tau}$ and every $b \in B_\tau$, one has $\left[J_\tau A_\tau J_\tau^{-1} x\right]_b = \iota_b^{-1}\left(A\left(\iota_{\pi_\tau^{-1}(b)}\left(x_{\pi_\tau^{-1}(b)}\right)\right)\right) = \left(\iota_b^{-1} \circ A|_{H_{\pi_\tau^{-1}(b)}} \circ \iota_{\pi_\tau^{-1}(b)}\right)\left(x_{\pi_\tau^{-1}(b)}\right) = g_b x_{\pi_\tau^{-1}(b)}$. Therefore, $J_\tau A_\tau J_\tau^{-1} = \widetilde{\rho}_\tau(g,\pi_\tau)$, and hence $A_\tau = \rho_\tau(g,\pi_\tau) \in \mathfrak{W}_\tau$. Since this holds for every type τ, one has $A \in \mathfrak{W}$. Combining $A \in \mathcal{G}_{\mathrm{glob}}$ and $A \in \mathfrak{W}$ gives $A \in \mathcal{G}_{\mathrm{exch}}$. Hence $\mathcal{G}_{\mathrm{act}} \leq \mathcal{G}_{\mathrm{exch}}$.

Assume first that $\mathcal{G}_{\mathrm{act}} = \mathcal{G}_{\mathrm{exch}}$. Every element of $\mathcal{G}_{\mathrm{act}}$ preserves $\mathfrak{F}_{\mathrm{adm}}$ and the observation map. Hence $\mathfrak{F}_{\mathrm{adm}}$ is invariant under $\mathcal{G}_{\mathrm{exch}}$, and the stated completeness condition holds.

Conversely, assume invariance and completeness. Let $A \in \mathcal{G}_{\mathrm{exch}}$. Invariance gives $A \cdot \mathfrak{F}_{\mathrm{adm}} = \mathfrak{F}_{\mathrm{adm}}$, and completeness gives $\Phi_{\mathcal{O}}(A \cdot f) = \Phi_{\mathcal{O}}(f)$ for every $f \in \mathfrak{F}_{\mathrm{adm}}$. Hence $A \in \mathcal{G}_{\mathrm{act}}$, so $\mathcal{G}_{\mathrm{exch}} \leq \mathcal{G}_{\mathrm{act}}$. Combining this with the upper bound proves equality. □

F Experimental protocols and additional results

F.1 Numerical settings and statistical reporting

The experiments used double-precision floating-point tensors. Integer certificates, token identifiers, indices, and Boolean masks retained their respective types. All experiments used Python 3.11.5, PyTorch 2.5.1 with CUDA 12.4, and an NVIDIA GeForce RTX 4090, with eight CPU threads, deterministic algorithms enabled, and TF32 disabled. The Pythia runs used Transformers 4.49.0, unquantized checkpoints, evaluation mode, batch size one, and no key/value cache. RoPE and eager-attention softmax were evaluated in FP64. Each model's first complete forward was audited at the tensor interface, recording 3582 FP64 tensor occurrences for Pythia-1.4B and 4678 for Pythia-2.8B; these are not counts of distinct operations or audits of subsequent forwards.

Seeds were $0, \ldots, 9$. Pythia means and sample standard deviations summarize seed-level averages over examples and the 16 fixed locations. Experiment 4 used one deterministic identity control. Error maxima use individual recorded checks. Relative Frobenius errors divide the difference norm by the baseline norm, with denominator floor 10^{-30}. Experiment 4 checked logits and contexts with elementwise absolute and relative tolerances of 10^{-8}.

F.2 Observation refinement

Experiment 1. For each activation matrix $Z \in \mathbb{R}^{d\times 16}$, we fixed the readout $U = [I_4\ 0]$. The coarse observation was $\Phi_{\mathrm{coarse}}(Z) = (UZ, \|Z\|_F^2)$; the refined observation was $\Phi_{\mathrm{refined}}(Z) =$

Table 3: Tests with no observed shear disagreement. Counts are out of 30 dimension–seed configurations at each setting. The analytic probability applies separately to either full-dimensional sampling distribution. Sample-size tests use nested prefixes.

Shear strength	Samples	Analytic miss (%)	Gaussian count	Ball count
0.001	64	97.98	28/30	30/30
0.001	256	92.17	25/30	28/30
0.001	1024	72.18	17/30	21/30
0.001	4096	27.14	6/30	10/30
0.01	64	81.54	25/30	27/30
0.01	256	44.21	11/30	17/30
0.01	1024	3.82	0/30	0/30
0.01	4096	2.13×10^{-4}	0/30	0/30

$(UZ, Z^\top Z)$. The binary assignment was $\beta(Z) = \mathbf{1}\{\mathrm{rank}(Z) > 5\}$; a constant assignment served as a control. For each of the three dimensions $d \in \{32, 64, 128\}$ and each seed, we generated 32 pairs. Their common leading four rows had a nonsingular integer diagonal block. One matrix had an additional entry $5t$, while the other used $3t$ and $4t$ in two separate rows and columns, with integer $t \in \{1, \ldots, 8\}$. Thus the readouts and squared norms agreed exactly, while nonzero triangular minors and the number of nonzero rows certified ranks 5 and 6.

All 960 pairs had identical coarse observations, different binary labels, and distinct refined observations; the refined Gram rank recovered every label. Rank preservation under invertible transformations certified that the two matrices in each pair belonged to different orbits. Four orthogonal transformations of the unobserved coordinates were applied to both matrices in each pair. These controls preserved the rank labels, with the largest Gram residual equal to 2.27×10^{-12}. The mean Frobenius distances between the distinct Gram observations were 641.1 ± 87.2, 575.4 ± 64.4, and 609.9 ± 97.9 for dimensions 32, 64, and 128, respectively; uncertainties use the ten means over 32 pairs. The intrinsic ranks and number of inputs were fixed across dimensions, so these results check the same low-rank construction in different ambient spaces.

F.3 Probe support and sample size

Experiment 2. In each dimension $d \in \{64, 256, 1024\}$ and seed, we constructed orthonormal vectors w, v, and c. Positive and negative controls used $A = I + (\chi - 1)ww^\top$, with $\chi = 1.7$ and $\chi = -1$. Shears used $A = I + \alpha wv^\top$, with $\alpha \in \{0.001, 0.01, 0.1, 1\}$, giving $A^\top w = w + \alpha v$. The four sampling distributions were a standard Gaussian in the ambient space, a uniform ball of radius 0.25 centered at c, a Gaussian projected onto $v^\perp$, and a uniform ball of radius 0.10 centered at $2w$. The first ball contains a point on the original decision boundary, whereas the second ball lies strictly on its positive side. Each sample-size test used a prefix of the same 4096 draws, at sizes 64, 256, 1024, and 4096; the prefixes are therefore not independent repetitions.

For the Gaussian and boundary-centered ball, the shear disagreement probability is $p(\alpha) = \arctan(\alpha)/\pi$, and the probability of observing no disagreement in n independent draws is $(1 - p(\alpha))^n$. Table 3 reports the observed counts of tests with no disagreement. At $n = 4096$, the Gaussian disagreement rates averaged over the 30 dimension–seed configurations were 0.0342%, 0.3247%, 3.1779%, and 24.9862% as the shear strength increased; the corresponding analytic rates were 0.0318%, 0.3183%, 3.1726%, and 25.0000%. Every transformation violating the positive-ray condition also had a directed ambient-space witness and a witness inside the boundary-centered ball. All sampled labels were preserved by every shear on the specified subspace and one-sided ball. Here "false acceptance" means using such agreement to accept the ambient-space positive-ray condition, not misclassifying whether the sampled labels agree.

F.4 Metric, readout, ReLU, and block controls

Experiment 3. The architectural experiments used $d \in \{64, 128, 256\}$, ten seeds, and 2048 sampled states per setting. For the metric/readout tests, $C = I$ and $U = [I_{r_U}\ 0]$, with $r_U \in \{0, d/4, d/2, 3d/4, d-1, d\}$. We constructed matrices acting as the identity on the observed coordinates and orthogonally on their complement, together with explicit reflections and scaling violations.

Table 4: Maximum absolute errors over the 30 dimension–seed configurations for each ReLU transformation family. The two commutation columns compare $\mathrm{ReLU}(Az)$ and $A\,\mathrm{ReLU}(z)$ on the specified states. The final column compares the compensated network with its original output.

Transformation	Mixed-sign states	Nonnegative states	Network output
Positive diagonal	0	0	2.44×10^{-15}
Permutation	0	0	2.66×10^{-15}
Positive monomial	0	0	3.55×10^{-15}
Positive nonmonomial shear	1.82	0	0.422
Orthogonal	4.22	5.49	3.69
Coordinate sign reversal	4.79	4.67	0.881

Their expected stabilizer has continuous dimension $(d - r_U)(d - r_U - 1)/2$. This dimension and the existence of a reflection were specified analytically by the construction, rather than estimated from numerical residuals. At $d = 128$, the dimensions for the six readout ranks were 8128, 4560, 2016, 496, 0, and 0. For every $r_U < d$, reflection in an unobserved coordinate preserved the metric and readout, reversed the associated probe, and changed an additional observation of that coordinate. The resulting witness shows directly that the metric/readout constraints need not preserve a richer observation protocol. At full readout rank, the constructed admissible transformation was the identity.

The ReLU controls compared positive diagonal matrices, permutations, positive monomial matrices, a positive nonmonomial shear, orthogonal matrices, and a single-coordinate sign reversal. For each seed, a two-layer ReLU network with 64 inputs, d hidden units, and 8 outputs supplied preactivations from 2048 Gaussian input vectors; coordinatewise absolute Gaussian vectors provided the nonnegative control states. We also transformed the network's first-layer weights and bias by A and compensated its second-layer weights by A^{-1}. Table 4 reports maxima across the dimensions and seeds. The shear $I + 0.5e_1e_2^\top$ had zero residual on nonnegative states, but its directed mixed-sign witness had residual 0.447214 in every configuration. Conversely, an orthogonal matrix could preserve the metric to FP64 precision while failing the ReLU constraint. The network check concerns these explicit two-layer constructions, not all symmetries of arbitrary multilayer ReLU networks.

The block experiments used four equal-size orthogonal blocks, with two blocks of each of two types. Within-block transformations satisfied both named-block and exchangeable-block constraints. Swapping two blocks of the same type violated the named projections but preserved the type structure; swapping blocks of different types violated both constraints. The specified within-block continuous dimensions were 480, 1984, and 8064 for the three ambient dimensions. The named-block permutation group contained only the identity, whereas the exchangeable same-type permutation group was $S_2 \times S_2$, with four elements. These counts follow from the specified architecture; the numerical tests check the corresponding projection and type-preservation conditions.

F.5 Pretrained checkpoints, data, and probes

The Pythia checkpoints were taken at training step 143000, with the following resolved model commits.

Pythia-1.4B	`9cc5c8c8148a4e0115d9e29c6b4f21124cfe748a`
Pythia-2.8B	`dbe7ae300a54abcdc475a33907b3dff81d25709f`

The models contained 1,414,647,808 and 2,775,208,960 parameters, respectively. Their FP64 parameter storage was 11,317,182,464 and 22,201,671,680 bytes; the smaller model had no CPU-resident parameters, while the larger model placed 1,258,823,680 parameter bytes on the CPU. The checkpoint configurations recorded FP16, so FP64 describes the experimental execution, not the precision of pretraining or the information retained in the stored checkpoint.

Pythia-1.4B used layers $\{0, 8, 15, 23\}$ and heads $\{0, 4, 8, 12\}$, with head dimension 128; Pythia-2.8B used layers $\{0, 10, 21, 31\}$ and heads $\{0, 8, 16, 24\}$, with head dimension 80. Indices are zero-based. All sixteen combinations were fixed before probe evaluation. Text was tokenized without added special tokens and truncated to at most 128 tokens. The feature at each location was the attention-weighted value vector at the final token, before multiplication by the head's output projection block.

Table 5: Fixed-probe prediction disagreement (%) on 872 cached representations, averaged over 16 layer/head locations and ten reparameterization seeds. ± denotes the SD across seed-level location means; positive and negative denote $A^\top w = 1.7w$ and $A^\top w = -w$.

Model	Positive	Negative	Orthogonal	General linear
Pythia-1.4B	0.00 ± 0.00	100.00 ± 0.00	51.79 ± 2.04	49.48 ± 4.38
Pythia-2.8B	0.00 ± 0.00	100.00 ± 0.00	50.42 ± 2.21	51.11 ± 2.57

We used the `stanfordnlp/sst2` data. Training texts were deduplicated after lowercasing and collapsing whitespace, and normalized exact matches to validation texts were excluded from the training pool. With data-selection seed 1729, we drew balanced, disjoint training and tuning sets of 2048 and 512 examples from the remaining training split. All 872 public labeled validation examples formed the evaluation set. This procedure removes normalized exact matches; it does not establish independence among related phrases. The full-forward verification subset contained eight examples of each label, fixed using seed 8832 and reused for every transformation.

Each bias-free logistic probe was fitted after coordinatewise root-mean-square scaling estimated on its training data, without centering; the fitted covector was converted back to the original feature coordinates. The objective was mean logistic loss with an ℓ_2 penalty on the scaled-coordinate coefficients. Regularization was selected from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ by tuning accuracy, breaking ties by lower tuning loss and then larger regularization. L-BFGS used at most 250 iterations, a strong-Wolfe line search, gradient tolerance 10^{-8}, and objective/parameter-change tolerance 10^{-12}. A shuffled-training-label control was fitted at the selected regularization for each location. Evaluation accuracy ranged from 48.62% to 83.03% in Pythia-1.4B and from 53.90% to 86.24% in Pythia-2.8B. These fixed locations do not give a monotonic comparison of probe quality across model sizes.

None of the 32 selected probes met the prescribed gradient tolerance. Final gradient infinity norms ranged from 8.10×10^{-8} to 3.39×10^{-5} in Pythia-1.4B and from 7.98×10^{-8} to 3.90×10^{-7} in Pythia-2.8B. Experiment 4 used these frozen, finite, nonzero covectors. The identities for $A^\top w$ and $A^{-\top} w$ apply to such covectors without assuming they minimize the probe objective exactly.

F.6 Transformations and complete-forward checks

Experiment 4. The transformation families were the identity, random orthogonal matrices, general linear matrices with condition number 3, and positive- and negative-ray controls. General linear matrices combined independent orthogonal factors with logarithmically spaced singular values from $3^{-1/2}$ to $3^{1/2}$. The ray controls imposed $A^\top w = 1.7w$ or $A^\top w = -w$, with random orthogonal action on the complement of w. Every nonidentity family used ten seeds and separate matrices at each fixed location; transported probes solved $A^\top w' = w$ without refitting.

Cached fixed-probe predictions. Positive-ray transformations preserved every fixed-probe prediction, while negative-ray transformations reversed every prediction (Table 5). Random orthogonal and well-conditioned general linear transformations changed approximately half of the fixed-probe predictions. For a sign-symmetric sampler with A and $-A$ equally distributed and no zero transformed scores, the expected disagreement is exactly one half (Proposition 8); this baseline is not a semantic-specific discovery. Covariantly transformed probes retained all original predictions. Mean baseline accuracies across the 16 locations were 68.65% and 64.13% for the two models, compared with 49.99% and 49.84% after shuffling the probe-training labels.

Cached evaluation covered all 872 examples. Complete-forward tests simultaneously changed value weights, value biases, and output-projection blocks at the 16 fixed locations, while leaving queries and keys unchanged. In exact arithmetic the transformed head context is Ah, while the compensated output is $W_O A^{-1} Ah = W_O h$. These are head-local value-space reparameterizations rather than changes of basis in the full residual stream. The 16-text verification compared all-token, all-vocabulary logits and measured contexts with their predicted transformed values.

Table 6 reports the numerical results. Covariantly transformed probes had zero disagreement in all complete-forward checks. Their largest absolute score errors were 6.79×10^{-12} and 6.73×10^{-13} for

Table 6: Actual-weight complete-forward verification. Disagreement is in percent, with sample standard deviation across ten seed means after averaging over the sixteen texts and sixteen locations. Error columns give the largest recorded relative Frobenius errors. Covariantly transformed probes had zero disagreement in all rows. Each model additionally had sixteen identity-control forwards, with zero errors.

Model	Transformation	Fixed disagreement	Logit error	Context error
1.4B	Positive ray	0.00 ± 0.00	5.12×10^{-14}	4.08×10^{-13}
1.4B	Negative ray	100.00 ± 0.00	5.09×10^{-14}	4.43×10^{-13}
1.4B	Orthogonal	52.03 ± 3.81	4.97×10^{-14}	4.79×10^{-13}
1.4B	General linear	49.84 ± 4.18	5.33×10^{-14}	5.63×10^{-13}
2.8B	Positive ray	0.00 ± 0.00	3.15×10^{-14}	1.09×10^{-13}
2.8B	Negative ray	100.00 ± 0.00	3.38×10^{-14}	1.35×10^{-13}
2.8B	Orthogonal	49.80 ± 3.70	3.16×10^{-14}	1.10×10^{-13}
2.8B	General linear	50.86 ± 3.94	3.40×10^{-14}	1.04×10^{-13}

Pythia-1.4B and Pythia-2.8B, respectively, while maximum absolute logit errors were 4.49×10^{-12} and 3.00×10^{-12}. These checks supplement the cached analysis on a fixed subset of 16 texts; they are not exhaustive input-domain tests.

Each model supplied 656 verification records, including 16 identity controls; all logit and internal-context checks passed. Saved parameter slices were restored after each intervention; a final reference-text forward reproduced logits and contexts bitwise in both models.

F.7 Finite calibration and pretrained interventions

Calibration matrices, transformations, and probe computations used FP64 with seeds $0, \ldots, 9$. Numerical rank counted singular values exceeding $32 \max(m, n)\epsilon_{64}\sigma_{\max}$ for an $m \times n$ matrix, where ϵ_{64} is machine epsilon and $\sigma_{\max}$ the largest singular value; empty and zero matrices had rank zero. Gaussian QR factorizations used diagonal sign correction. Householder reflections were replaced by identity when the uncalibrated probe component had norm at most 10^{-12}.

Experiment 5. Support dimensions were $r_L \in \{16, 32, 64\}$, ambient settings were $d \in \{64, 256, 1024\}$, and calibration sizes were $n_{\mathrm{cal}} \in \{0, r_L/4, r_L/2, 3r_L/4, r_L - 1, r_L\}$. Computation remained in support coordinates; ambient embeddings were checked for orthonormality but not used in evaluation. Calibration matrices contained independent Gaussian columns or cyclic repetitions of $\lceil n_{\mathrm{cal}}/2 \rceil$ Gaussian columns. A third control selected the probe as the leading left singular vector of a nonempty Gaussian calibration matrix; other probes were normalized Gaussian vectors. Identity, Householder reflections, random orthogonal transformations fixing the calibration span, and unconstrained orthogonal transformations each used 65,536 Gaussian evaluation states in batches of 8192.

For the beta-law test, each seed fixed a probe before drawing 100 independent calibration matrices per interior calibration size. The 1000 values per setting were pooled for twelve Kolmogorov–Smirnov tests. For random calibration of intermediate rank, boundary states summed the span and orthogonal-complement probe components divided by their respective squared norms, provided both norms exceeded 10^{-12}. Their original and reflected scores are respectively two and zero in exact arithmetic.

Experiment 6. Checkpoints, locations, feature definitions, tokenization, and probe-fitting procedures followed Appendix F.5. Training and tuning sets were drawn separately from Experiment 4, with a disjoint balanced calibration pool of 256 training texts, using seed 1729. Calibration sizes were $\{0, 8, 16, 32, 64, d_h - 1, d_h\}$, where d_h is the head dimension. Subsets were sampled without replacement, shared across locations, and not nested across sizes. Probes remained fixed across interventions.

Cached evaluation used all four transformation families on 872 validation texts. Covariant probes solved $A^\top w' = w$ without refitting. Complete-forward evaluation excluded identity and used sixteen balanced validation texts selected with seed 8832, giving 420 model–size–seed–family config-

urations. Value weights, biases, and output blocks at all sixteen locations were modified simultaneously; orthogonal output compensation used $A^\top$. Pythia parameters and rotary computations used FP64; the tensor-operation audit described above applies to Experiment 4.

Calibration residuals compared transformed cached calibration vectors with their originals, rather than recomputing calibration-text forwards. Complete-forward checks compared all-token logits and last-token contexts with their predicted transformations. Errors were recorded without tolerance-based rejection; parameter restoration was checked exactly. Disagreement means and sample standard deviations were computed across seed-level averages over examples and locations.

Pythia calibration results. Separate calibration runs used the same checkpoints, locations, and probe-fitting protocol, with mean baseline accuracies of 68.16% and 65.01%. Calibration contexts came from a disjoint pool of 256 SST-2 training texts; evaluation used all 872 validation examples. With $n_{\mathrm{cal}} = 8$, the Householder reflection changed $65.96 \pm 3.13\%$ and $68.57 \pm 3.26\%$ of fixed-probe predictions in Pythia-1.4B and Pythia-2.8B, respectively. At $n_{\mathrm{cal}} = 64$, these rates fell to $36.00 \pm 1.54\%$ and $21.97 \pm 1.01\%$, compared with $23.71 \pm 0.99\%$ and $15.39 \pm 0.81\%$ for random calibration-preserving transformations. Unconstrained orthogonal controls remained near 50%. The largest relative Frobenius residual on the cached calibration contexts for the constrained transformations was 1.05×10^{-12}, and covariantly transformed probes had zero disagreement. At 128 calibration states, Pythia-1.4B had zero constrained disagreement; at 80, Pythia-2.8B retained $0.13 \pm 0.22\%$ reflection disagreement, with four of 160 location–seed matrices having numerical rank below 80.

Repeated and layerwise checks. The 420 joint-intervention configurations modified all sixteen target locations simultaneously and were evaluated on sixteen fixed verification texts. Of these configurations, 419 had maximum relative Frobenius logit error at most 4.21×10^{-15}. The remaining Pythia-2.8B reflection, with calibration size eight and seed five, reached 2.013×10^{-6} on validation example 426, with maximum absolute logit error 1.180×10^{-4}. Two fresh model reloads reproduced the discrepancy. On this example, diagnostics evaluated sixteen single-location, four single-layer, and four cumulative-layer scenarios. They recorded logits, head contexts, pre-layer and post-layer residual states, and parameter-restoration errors, comparing residuals over both the full sequence and the final token. Every single-location and single-layer intervention had relative Frobenius logit error at most 1.81×10^{-15}, whereas joint interventions in layers 0 and 10 reproduced the discrepancy. For that joint intervention, the full-sequence relative residual error increased from 5.03×10^{-16} to 6.33×10^{-11} across the unmodified layer 14. Covariant predictions remained unchanged, and restored parameter slices had zero error. These observations are consistent with finite-precision error amplification, but do not identify the responsible operation within that layer.

F.8 GPT-2 ARCHITECTURAL INTERVENTIONS

Model. Experiment 7 used the complete pretrained `openai-community/gpt2` checkpoint at commit `607a30d783dfa663caf39e06633721c8d4cfcd7e`. It has 12 layers, 12 attention heads per layer, residual width 768, head width 64, and 124,439,808 unique parameters. The native token-embedding and language-model output weights remained tied. We used the stored `gelu_new` activation and LayerNorm epsilon 10^{-5}, without quantization. Execution used Python 3.11.5, PyTorch 2.5.1 with CUDA 12.4, Transformers 4.49.0, and an NVIDIA RTX 4090, with eight CPU threads. Parameters, floating buffers, and numerical computations used FP64, deterministic algorithms were enabled, and TF32 was disabled. Evaluation used eager attention, batch size one, no key/value cache, and disabled dropout. FP64 refers to experimental execution, not pretraining precision.

Data. The SST-2 revision was `8d51e7e4887a4caaa95b3fbebbf53c0490b58bbb`. The dataset identifier was `stanfordnlp/sst2`. After lowercasing and collapsing whitespace for duplicate detection, training duplicates and normalized exact matches to validation texts were excluded. Seed 1729 selected a balanced training set of 2048 examples and a disjoint balanced calibration pool of 256 examples. Tokenization added no special tokens and truncated texts to 96 tokens. Cached evaluation used all 872 labeled validation examples. The residual feature was the last-token output of each Transformer block, before final LayerNorm at the last block. The head feature was the last-token attention-weighted value vector before the head output projection. All 12 residual locations and 144 head locations were included, without selecting locations by accuracy.

Probe fitting. Each location used a bias-free logistic probe with training-set coordinatewise root-mean-square scaling and no centering. Writing z_i for the scaled features and $y_i \in \{-1, 1\}$, we minimized $n^{-1} \sum_i \log(1+\exp(-y_i z_i^\top \theta)) + 0.01\|\theta\|_2^2/2$. The regularization coefficient was fixed rather than tuned. Damped Newton optimization used at most 100 iterations, at most 40 line-search steps, Armijo coefficient 10^{-4}, and gradient infinity-norm tolerance 10^{-9}. RMS scales had floor 10^{-12}; fitted coefficients were converted back to the original feature coordinates. A shuffled-training-label probe was fitted independently at every location using the same regularization. All 156 fitted probes and all shuffled controls met the gradient tolerance. Their largest gradient infinity norms were 8.46×10^{-10} and 9.71×10^{-10}, respectively. Mean validation accuracies were 79.03% for residual probes and 67.63% for head probes, compared with 49.84% and 49.67% for shuffled controls. These are location averages for one checkpoint and fixed probes, not averages over independently trained models.

Attention reparameterizations. We use row-vector notation in this subsection. Query–key and value–output compensations use the known attention symmetries (Tran et al., 2025; Silva et al., 2025). For every head, $Q' = QS$, $K' = KS^{-\top}$, $V' = VB$, and $W'_O = B^{-1}W_O$ preserve $QK^\top$ and the head output, with the corresponding projection biases transformed as well. Matrices $S, B \in \mathrm{GL}(64)$ had condition number two, constructed from independent Haar orthogonal factors and logarithmically spaced singular values. Head permutations moved the query, key, value, bias, and output blocks together. They preserve model outputs but not named head identities. Fixed-head comparisons retained the physical head index; transported-probe comparisons followed the source-head permutation. Thus zero transported disagreement under a head permutation refers to the corresponding original head, not the unchanged physical index.

Global residual transformations. For a permutation matrix P and common sign $s \in \{-1, 1\}$, set $T = sP$ and transform residual rows by $h' = hT$. Token and position embeddings were multiplied by T. Each residual-input projection used $W' = T^\top W$, while each residual-output projection and its bias used $W' = WT$ and $b' = bT$. Every LayerNorm, including the final one, used $\gamma' = \gamma P$ and $\beta' = \beta T$. These transformations give $\mathrm{LN}(hT; \gamma P, \beta T) = \mathrm{LN}(h; \gamma, \beta)T$ and preserve the tied output readout. Residual negation used $T = -I$, whereas residual permutation used $T = P$. The joint condition composed a negated residual permutation with all-head query–key and value–output compensations, head permutations, MLP permutations, and pre-LayerNorm rescaling. These are explicit admissible transformations, not a search over all functionally equivalent networks.

MLP, normalization, and affine controls. MLP permutations reordered the first projection's output coordinates and bias and the second projection's input coordinates in every layer. Pre-LayerNorm rescaling multiplied each affine output coordinate by a nonzero signed log-normal scale and inversely scaled its consuming projection. Three additional conditions tested affine symmetries. Key-bias translations exploit softmax row-shift invariance. Value-bias translations were compensated in the attention output bias. Residual mean translations added one scalar multiple of the all-ones vector to each MLP output bias; subsequent LayerNorm removes the accumulated mean shift. Translation samples had scale 0.1. These affine conditions are separate from the homogeneous linear-probe results. For $h' = hT + t$, score preservation uses $w' = T^{-1}w$ and intercept $c' = -tw'$, so that $h'w' + c' = hw$. No probe was refitted after an intervention.

Counterfactual parameter changes. Five controls deliberately violated the compensation requirements. The query–key control used S^{-1} rather than $S^{-\top}$ for the keys. The value control omitted output compensation. The GELU control multiplied every MLP preactivation by 1.7 and divided the following projection by 1.7. Another control permuted residual coordinates but left LayerNorm parameters unchanged. Finally, a dense residual candidate used a Householder reflection R fixing the all-ones vector. With each pre-LayerNorm gain and bias fixed, its consuming projection was changed to $W' = D_\gamma^{-1} R^\top D_\gamma W$ and $b' = b + \beta W - \beta W'$, where $D_\gamma = \mathrm{diag}(\gamma)$. Residual-output projections and embeddings were multiplied by R. The final gain remained fixed, its bias became $\beta' = \beta D_\gamma^{-1} R D_\gamma$, and native weight tying was retained. This correctly compensates the block inputs but does not generally compensate the final normalization and tied readout. Failure of this candidate does not rule out other parameter transformations.

Table 7: GPT-2 counterfactual controls on the same 32 validation texts. Logit errors are relative Frobenius errors. Argmax disagreement is the percentage of token positions with changed highest-logit tokens, averaged within text and then within seed. Uncertainties are sample standard deviations across seed means.

Intervention	Mean error	Maximum error	Argmax disagreement
Incorrect query–key inverse	0.557 ± 0.082	1.064	88.05 ± 1.56
Uncompensated values	0.634 ± 0.067	0.923	92.41 ± 3.07
GELU scaling	0.337 ± 0.000	0.467	66.35 ± 0.00
Dense residual candidate	0.081 ± 0.065	0.220	84.85 ± 12.79
Residual permutation without LayerNorm compensation	0.930 ± 0.097	1.107	97.71 ± 1.70

Specified stabilizer and rank diagnostics. For the stored final LayerNorm gain, we computed the restricted group $K_\gamma = \{R \in O(768) : R\mathbf{1} = \mathbf{1},\ RD_\gamma = D_\gamma R\}$. Grouping exactly equal stored gain values gives eigenspace multiplicities m_j. Commutation preserves these eigenspaces, and fixing $\mathbf{1}$ fixes its component in each one, giving $K_\gamma \simeq \prod_j O(m_j - 1)$. All 768 stored gain values were distinct, so $K_\gamma = \{I\}$. This is exact for the stated matrix constraints on the stored gains. It is not the full residual-stream symmetry group; the residual permutations above also transform LayerNorm parameters. Separately, numerical ranks counted singular values exceeding $32 \max(m, n)\epsilon_{64}\sigma_{\max}$ for an $m \times n$ matrix. All query, key, value, and head-output matrices had numerical rank 64, and the tied embedding had numerical rank 768. These diagnostics establish neither exact algebraic rank nor headwise controllability. The parameter dimensions of the specified transformation families were recorded separately and were not summed into a claimed complete-group or orbit dimension.

Complete-forward evaluation and aggregation. The 20 conditions comprised the twelve conditions in Table 1, three calibration sizes in Table 2, and the five controls in Table 7. Seeds $0, \ldots, 9$ varied transformations or calibration selection, with the checkpoint, probes, and evaluation inputs fixed. Identity, residual negation, and the fixed GELU scaling were deterministic conditions repeated across seeds. Each configuration used the same 32 balanced validation texts, selected with seed 8832, and eight fixed random-token sequences, two each of lengths 8, 16, 32, and 64. After actual weight modification, we compared all-token, all-vocabulary logits, layer residuals, head contexts, and attention probabilities. For logits Z, relative error was $E_{\text{logit}} = \|Z' - Z\|_F / \max(\|Z\|_F, 10^{-30})$. Passing required both $E_{\text{logit}} \leq 10^{-9}$ and maximum absolute error at most 10^{-7}. Internal comparisons required relative error at most 10^{-8} and maximum absolute error at most 10^{-7}. Logit errors were averaged over texts within each seed; probe disagreements additionally averaged over the fixed locations. Reported standard deviations use these ten seed means. Maxima refer to individual recorded comparisons, not seed means.

Finite calibration. For each seed and calibration size $n_{\text{cal}} \in \{8, 32, 64\}$, contexts were selected without replacement from the disjoint pool. The same texts were used at all heads; subsets were not nested across sizes. At each head, let $v = w - P_{\text{cal}} w$, where the calibration-span projection used the stated SVD threshold. We applied the Householder matrix $B = I - 2vv^\top / \|v\|_2^2$ with compensated value and output projections in all 144 heads. It was replaced by identity when $\|v\|_2 \leq 10^{-12} \max(\|w\|_2, 1)$. Every selected calibration text was recomputed through the modified complete model, in addition to the fixed verification set. Calibration-context error compares the layer's matrix of all last-token head contexts with its original value. Unlike the Pythia checks in Experiment 6, these calibration errors were measured after recomputing the calibration texts. Table 2 separates the full validation-set cached analysis from actual forwards on the 32 verification texts. All transported-probe disagreements were zero. At calibration size 64, recorded constrained disagreements and forward errors were zero; this is not a population-support certificate. Neither the Gaussian optimum nor the beta law is asserted for these pretrained features.

Numerical checks and interpretation. Across all compensated conditions, including the calibration, validation, and random-token inputs, the maximum relative Frobenius logit error was 6.34×10^{-15}; logit and internal checks passed. Fixed residual-probe predictions reversed under residual negation despite zero recorded logit error. LayerNorm compensation distinguished the

valid residual permutation from its counterfactual, whose mean relative logit error was 0.930. The run recorded $12{,}510$ tensor-interface forward audits with no non-FP64 floating tensors, comprising 3176 feature-extraction forwards without the language-model head and 9334 complete-model forwards. The latter included 9040 intervention forwards, 293 baseline forwards, and one final restoration check. All unique parameters were restored and compared exactly after each intervention. The final reference-text check reproduced logits, residuals, contexts, and attention probabilities bitwise. These audits concern observable tensor dtypes, not arithmetic inside device kernels. The experiment concerns the specified reparameterizations and fixed inputs, without claiming a complete characterization of observation-preserving transformations, orbit separation, or population support.