THE LINEAR REPRESENTATION HYPOTHESIS NEEDS A GROUP ACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

To make claims about representations that generalize beyond a particular trained model, we need to specify when two representations should count as equivalent. The Linear Representation Hypothesis is often discussed without making this equivalence explicit. Different notions of equivalence preserve different structures, so metrics, probes, and interventions that appear to study the same representation may in fact correspond to different hypotheses. We therefore argue that the Linear Representation Hypothesis is not one hypothesis but a family of claims distinguished by representation equivalence. We formalize this idea using group actions, specifying the representation object, the procedure that produces it, and the property ultimately asserted, while accounting for equivalences imposed by the model architecture. This framework clarifies how assumptions can change across metrics, reading points, and analysis stages, and we use it to audit common representation quantities and recent interpretability analyses.

1 INTRODUCTION

The Linear Representation Hypothesis underlies much of modern interpretability, whether it is tested explicitly (Park et al., 2024; Marks & Tegmark, 2024) or assumed when constructing linear probes (Alain & Bengio, 2016; Gurnee & Tegmark, 2024; Yao et al., 2026), steering interventions (Rimsky et al., 2024; Liu et al., 2026; Raval et al., 2026), learned dictionaries (Bricken et al., 2023; Templeton et al., 2024), and methods for comparing representations (Kornblith et al., 2019; Williams et al., 2021; Domenichelli et al., 2026). Such studies usually aim to draw conclusions about representations rather than artifacts of a particular realization, which requires specifying when two realizations count as the same representation. Calling a representation "linear" does not answer this question: one must also specify which changes are merely changes of description.

Existing formulations rarely make this equivalence explicit. Different analyses implicitly choose different answers, so methods that appear to study the same representation may in fact be probing different objects and testing different hypotheses. The issue can arise even within a single study.

Arditi et al. (2024), for example, estimate a difference-in-means vector and use the same array in two ways. It is added to activations without normalization, where it acts as a displacement in V and its magnitude affects the intervention, and separately normalized before its span is projected out, where only its projective direction in $\mathbb{P}(V)$ matters. Both interventions are effective and are reported as evidence for a single refusal direction, even though the two uses place that direction in different mathematical spaces and require different notions of equivalence.

Neither operation is invalid, since each is well defined under its assumed structure. The problem is that these assumptions are often unstated and may change between estimation, comparison, and intervention. As a result, different stages of an analysis can silently refer to different notions of the representation, while the final conclusion is stated as though they referred to the same object.

We propose that a representation claim be specified by the space M in which its object lives together with the action of an equivalence group G, the procedure F that produces the object, and the predicate P ultimately asserted of it. The procedure must transform equivariantly under changes of representation, while the predicate must remain invariant. The choice of G is also constrained by function-preserving reparameterizations of the architecture. If a function-preserving reparameteri-

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.

zation changes the value of a quantity computed from an internal representation, that quantity is a property of the parameterization rather than of the model.

Using this specification, we show that the Linear Representation Hypothesis is not one hypothesis but a family of claims. Displacement, decoding, superposition, and subspace formulations place features in different mathematical objects and admit different transformation laws. The analogous distinction applies to the procedures used to estimate these objects: two procedures may produce objects of the same type while respecting different equivalences. We further show that admissible symmetry depends on where a representation is read and that the symmetry of a multi-stage analysis must be checked for the composition as a whole.

Our position is that this specification is part of stating a representation claim rather than metadata attached to it. Making it explicit may narrow some conclusions, but it makes clear when two analyses are genuinely studying the same representation hypothesis.

The paper proceeds as follows. Sections 2–3 motivate and formalize representation equivalence. Section 4 develops its architectural and compositional consequences. Sections 5–6 audit common quantities and recent interpretability analyses. Section 7 extends the argument to the Platonic Representation Hypothesis, and Section 8 states the reporting principle.

2 The Linear Representation Hypothesis Is Not One Hypothesis

The Linear Representation Hypothesis is not a single claim. Different formulations place features in different mathematical objects and admit different transformation laws under which those objects remain meaningful. We first review some common formulations below.[1]

A classical formulation treats a feature or relation as a *displacement*. In word embeddings, some lexical relations were found to satisfy approximately constant offsets, $h_b - h_a \approx v$ (Mikolov et al., 2013). Modern methods such as difference-in-means (Marks & Tegmark, 2024) and contrastive activation addition (Rimsky et al., 2024) use the same basic object. Under an affine transformation $h' = Ah + t$, the displacement transforms as $v' = Av$, since the translation cancels. The coordinates of the vector change, but the constant-offset relation does not. This formulation therefore places the feature in V, with the primal affine action $v \mapsto Av$.

Linear decodability provides a different formulation. If a property can be recovered from an activation $h \in V$ by a linear functional $w \in V^*$ through $w^\top h$ (Alain & Bengio, 2016), then an invertible transformation $h' = Ah$ is accompanied by the dual action $w' = A^{-\top} w$, which preserves the readout. With a bias term, $w^\top h + b$, this extends to affine transformations by taking $b' = b - w'^\top t$. Thus even within probing, the admissible group depends on whether the probe includes a bias term. More importantly, the object has changed: a probe lives in V^* rather than V. Although probe weights and steering vectors are both stored as d-dimensional arrays, they obey different transformation laws and cannot be canonically identified without additional structure. Related work similarly distinguishes measurement from intervention geometry and linear representation from linear decodability (Park et al., 2024; 2025b; Garg et al., 2026).

Superposition gives another meaning to linear representation. In this formulation, an activation is written as $h = b + \sum_i x_i W_i$, where W_i denotes a feature direction and x_i its coefficient (Elhage et al., 2022). This additive view later motivated dictionary-learning approaches and sparse autoencoders (Bricken et al., 2023; Templeton et al., 2024). As a claim about the activation space, the decomposition is affine-covariant: taking $b \mapsto Ab + t$ and $W_i \mapsto AW_i$ reproduces the transformed activation with the coefficients unchanged. It also has a distinct ambiguity in its latent coordinates, reduced but generally not eliminated by sparsity or other constraints (Section 5). This latent non-identifiability is distinct from equivalence under changes of activation coordinates.

These examples show that "linear representation" does not determine a unique object or transformation law. Moreover, even for a fixed object space, different procedures can impose different symmetry requirements. We separate these choices before formalizing them in Section 3.

The object itself must be specified. Linearity also need not imply that a feature is one-dimensional (Olah, 2024). In our framework, a one-dimensional feature naturally lives in $\mathbb{P}(V)$, while a k-

[1] Notation and basic mathematical preliminaries are summarized in Appendix A.

dimensional linear feature lives in $\mathrm{Gr}(k, V)$. Circular representations of days and months provide an example in which a two-dimensional structure mediates computation and no single direction suffices (Engels et al., 2025). This distinction is compatible with linearity because the object space and admissible transformation group are separate parts of the claim: an invertible linear map preserves both lines and k-dimensional subspaces.

The object alone also does not determine the symmetry of the claim. A difference-in-means direction and a leading principal direction can both lie in $\mathbb{P}(V)$, but the procedures that produce them have different symmetry requirements. Difference-in-means requires no inner product, whereas PCA requires one to order directions by explained variance. Thus two procedures may return objects in the same space while being equivariant under different groups.

This distinction has an immediate consequence for empirical evidence. None of the formulations above intrinsically requires an inner product. Yet common evidence for them does: cosines and spectral quantities require similarity geometry, while absolute norms and intervention magnitudes require a fixed scale. The metric therefore enters through the evidence used to support the hypothesis rather than through the hypothesis itself.

This perspective differs from work that fixes a nuisance group to define representation similarity. In the generalized shape metrics of Williams et al. (2021), for example, quotienting by a chosen group yields a metric on representation space. Our position differs. The group is presupposed by the claim rather than chosen to make a measurement well behaved, so it is a property of the hypothesis. And it also constrained by the architectural floor and by the analysis pipeline (Sections 4.1 and 4.2).

3 Specifying a Member of the Family

Section 2 showed that formulations of the Linear Representation Hypothesis must specify not only what object represents a feature, but also how that object transforms and how it is obtained. We now formalize these choices as a specification of a representation claim.

3.1 Representation Equivalence

Let $\mathcal{X}$ be an input domain, V a finite-dimensional vector space, and $\mathcal{S} = \{\phi : \mathcal{X} \to V\}$ the space of representations. Suppose a group G acts on $\mathcal{S}$. In the settings considered below, this action is pointwise on the representation values in V: for each $g \in G$, there is a transformation $T_g : V \to V$ such that

$$(g \cdot \phi)(x) = T_g(\phi(x)). \tag{1}$$

This action specifies which transformations count as changes of description rather than changes in the underlying representation.

Definition 3.1 (Representation equivalence)**.** Two representations $\phi_1, \phi_2 \in \mathcal{S}$ are equivalent under G, written $\phi_1 \sim_G \phi_2$, if $\phi_2 = g \cdot \phi_1$ for some $g \in G$. The equivalence class of ϕ is its orbit $[\phi]_G = \{g \cdot \phi : g \in G\}$.

The choice of equivalence group determines the strength of the representation claim. A larger group leaves fewer quantities invariant, while a smaller group risks promoting coordinate-specific properties to properties of the representation. The equivalence relation is therefore part of the substantive claim, and it is not always freely chosen. A function-preserving reparameterization may alter the coordinates of an internal representation, so a claim about the model rather than one parameterization must remain valid across such realizations. The admissible group must therefore contain the transformations the architecture already realizes, a constraint made precise in Section 4.

3.2 Common Choices of Transformation Group

We now specialize the pointwise transformations in Equation (1) to affine maps $T_g(h) = Ah + t$, where A and t may depend on g. Within this class, three common transformation groups form the hierarchy

$$G_{\mathrm{iso}} \subset G_{\mathrm{sim}} \subset G_{\mathrm{aff}}. \tag{2}$$

The affine group G_{aff} allows arbitrary invertible A and translations t. The similarity group G_{sim} restricts the linear part to $A = sQ$, where $s > 0$ and Q is orthogonal. And the isometry group G_{iso} further requires $s = 1$.

These groups preserve progressively stronger structures. Affine transformations preserve affine relations, collinearity, subspace dimension, and intersection structure, but not angles or lengths. Similarities additionally preserve angles, orthogonality, and cosine similarity. Isometries further preserve norms, distances, and fixing absolute scale. A larger equivalence group therefore imposes stronger invariance requirements, while allowing fewer quantities to be attributed to the representation.

3.3 The Object and the Procedure

The equivalence group does not fully specify a representation claim. An analysis must also specify what mathematical object is extracted from the representation and how that object is constructed. In practice, the construction uses finitely many sampled activations $\{\phi(x_i)\}_{i=1}^n$. Under the pointwise actions of Section 3.2, these samples transform with the same coordinate change.

Let M denote the space of mathematical objects, equipped with an action of G, and let $F : \mathcal{S} \to M$ denote the procedure that constructs the object. A claim is a predicate P on M, so the resulting statement about ϕ is $P(F(\phi))$. For the statement to be independent of the chosen representation coordinates, the procedure and predicate must satisfy

$$F(g \cdot \phi) = g \cdot F(\phi), \qquad P(g \cdot m) = P(m). \tag{3}$$

It then follows that $P(F(g \cdot \phi)) = P(g \cdot F(\phi)) = P(F(\phi))$. Notice that the condition Equation (3) is stronger than invariance of the composite predicate $P \circ F$. We impose this factorization because the intermediate object is itself part of the representation claim and may be used in subsequent analyses. Its transformation law is therefore substantive: specifying M requires specifying not only what kind of object it contains, but also how G acts on that object.

This distinction already matters for two of the most common objects in representation analysis. A displacement is naturally an element of V and transforms as $v \mapsto Av$, whereas a linear probe is naturally an element of V^* and transforms as $w \mapsto A^{-\top} w$. Although both are vectors in the implementation, they belong to different G-spaces. The following proposition makes precise what additional structure is required to identify them.

Proposition 3.2 (Primal and dual objects)**.** *Let* $\dim V \geq 2$ *and let* $\mathrm{GL}(V)$ *act on* V *by* $v \mapsto Av$ *and on* V^* *by* $w \mapsto A^{-\top} w$*. Then:*

i. There is no nonzero $\mathrm{GL}(V)$*-equivariant map* $V^* \to V$*.*
ii. An inner product g *induces an equivariant map* $\sharp_g : V^* \to V$ *under* G_{iso}*.*
iii. The induced map $[w] \mapsto [\sharp_g w]$ *on projective spaces is equivariant under* G_{sim}*.*

A metric thus supplies an identification between V^* and V only under isometries, while passing to projective directions removes sensitivity to uniform scale and enlarges the symmetry to similarities. Regularized probe fitting provides another example: the ℓ_2 penalty breaks equivariance under general $\mathrm{GL}(V)$ transformations, leaving only isometric equivariance.

3.4 The Specification

The preceding components can now be collected into a complete specification:

Complete specification requires G, M, F, and P.

Definition 3.3 (Representation claim)**.** Let group G act on space $\mathcal{S}$. A representation claim consists of a G-space M, a map $F : \mathcal{S} \to M$, and a predicate P on M. The claim about $\phi \in \mathcal{S}$ is $P(F(\phi))$.

The G-space specifies both the mathematical object and how it transforms. The map F is separate because the same object space may be reached by procedures with different symmetry properties.

Definition 3.4 (Admissibility)**.** A representation claim (G, M, F, P) is admissible under G if

(A1) F is equivariant and P is invariant under the declared actions, as in Equation (3).
(A2) The action of G on $\mathcal{S}$ contains the architecture-induced equivalences described in Section 4.

Condition (A1) is internal to the analysis: it requires the method and conclusion to be well defined on the declared equivalence classes rather than on a selected coordinate realization. Condition (A2) supplies an external floor. If the architecture realizes a function-preserving transformation, a claim about the model cannot distinguish representations related by it. An analysis may therefore choose a larger group, but not a smaller one than the architecture permits.

Definition 3.5 (Comparability)**.** Two admissible claims are directly comparable if they use the same representation equivalence and the same G-space. Claims on different G-spaces require an explicit equivariant map relating those spaces and are otherwise not directly comparable.

Non-comparable claims may both be correct without supporting one another: equal numerical shape is insufficient unless the objects belong to spaces carrying compatible actions. When the object space is shared and $G_1 \subseteq G_2$ with the G_1 action obtained by restriction, (A1) under G_2 implies (A1) under G_1, although (A2) must be rechecked. Restricting the group can therefore make additional predicates well defined, as when an object estimated under a larger group is later reported through angles or norms requiring a smaller one.

4 Two Consequences of the Specification

The specification has two immediate consequences: an external constraint imposed by the architecture (A2), and an internal constraint arising from composition (A1).

4.1 The Architectural Floor

The equivalence group is not always freely chosen. Some transformations arise from the architecture itself at the parameter level and constrain which equivalence groups are admissible. Let Θ be the parameter space, let f_θ denote the function computed by parameters θ, and define the group of function-preserving parameter transformations as

$$\Gamma = \{\gamma : \Theta \to \Theta \mid \gamma \text{ invertible and } f_{\gamma\cdot\theta}(x) = f_\theta(x) \text{ for all } \theta, x\}. \tag{4}$$

At a fixed reading point, a parameter symmetry induces a representation action when the transformation descends to the activations. Writing $\phi^\theta : \mathcal{X} \to V$ for the representation at a fixed reading point, such an action exists when $\phi^{\gamma\cdot\theta}(x) = \rho(\gamma)\phi^\theta(x)$ for all $x \in \mathcal{X}$. The image of ρ is the architectural symmetry group at that reading point.

Proposition 4.1 (Architectural invariance)**.** *Suppose $\gamma \in \Gamma$ induces $\rho(\gamma)$ at a reading point, and let (G, M, F, P) be a representation claim satisfying (A1). If $P(F(\phi^{\gamma\cdot\theta})) \neq P(F(\phi^\theta))$ for some $\theta \in \Theta$, then the claim is not a property of the model function: it takes different values on parameter settings that compute the same function.*

Thus condition (A2) requires $G \supseteq \mathrm{Im}\rho$ at the reading point.

Corollary 4.2 (Reading-point dependence)**.** *The same architecture can induce different symmetries at different reading points. In dot-product attention, $W_Q \mapsto \Lambda W_Q$ and $W_K \mapsto \Lambda^{-\top} W_K$ preserve the model function for any $\Lambda \in \mathrm{GL}(d_{\mathrm{head}})$. At the residual stream this transformation acts trivially, so it excludes no angular claim there. At a query or key site, however, $\mathrm{Im}\,\rho$ contains $\mathrm{GL}(d_{\mathrm{head}})$, which admits no nonzero invariant bilinear form: taking $\Lambda = cI$ gives $b(cv, cw) = c^2 b(v, w) = b(v, w)$ for all $c > 0$, forcing $b = 0$. Angular and metric predicates defined from a fixed inner product are therefore not invariant at those sites.*

Architecture-induced symmetries of internal representations have been studied previously (Godfrey et al., 2022), including joint query–key rotations in transformers (Zhang et al., 2025). With RoPE, the induced symmetry is reduced but remains generally anisotropic, so the same conclusion holds. The derivation is given in Appendix C.

Equal-dimensional representations at different reading points may carry different G-actions, so transporting a direction between them requires an explicit equivariant map in the sense of Definition 3.5, rather than identification by shared coordinates.

4.2 Composition of Analysis Stages

Representation analyses are usually pipelines rather than single maps. Therefore, admissibility must be established for the composite procedure rather than inferred from individual stages.

Proposition 4.3 (Composition). *Let $F = F_n \circ \cdots \circ F_1$ with $F_i : M_{i-1} \to M_i$. If G acts on every M_i, each F_i is equivariant under the corresponding actions, and P is invariant on M_n, then (G, M_n, F, P) satisfies (A1) under G.*

If a stage is not equivariant, admissibility must instead be established for the composite. Its symmetry is not generally obtained by intersecting groups assigned to the stages in isolation, because a stage may change the object space and hence the action seen by the next stage. For example, under $h \mapsto Ah + t$, centered activations transform as $h - \bar{h} \mapsto A(h - \bar{h})$, so cosine after centering can be invariant to translations that would change cosine on the uncentered activations.

When all stages carry compatible actions of a common group, the pipeline is limited by its most restrictive stage. A direction may be estimated under G_{aff}, compared by cosine under G_{sim}, and calibrated by a norm under G_{iso}, so the resulting claim is guaranteed admissible only under G_{iso} unless stronger invariance of the composite is established.

5 A Symmetry Audit of Common Representation Quantities

Table 1 records, for quantities in common use, the largest group within the hierarchy considered here under which each is well defined; using a smaller group narrows the claim it can support.

Table 1: Symmetry audit of common representation quantities. Rows are grouped by the largest group under which the quantity is preserved, generically in the spectral parameters. The final column names the structure that fixes the restriction.

Quantity	Computed from	What fixes the group
Preserved under G_{aff}		
Exact rank	Centered activations	Linear dependence only
Span	Centered activations	Linear dependence only
Containment, intersection dim.	Subspace pair	Incidence structure
CCA	Representation pair	Centered linear relations
Preserved under G_{sim}		
Leading principal subspace	Covariance form	Variance ordering
Relative-threshold rank	Covariance spectrum	Spectral ratios
Effective rank, participation ratio	Covariance spectrum	Spectral ratios
Intrinsic dimension	Neighbour distances	Distance ratios
Cosine, angle, orthogonality	Direction pair	Inner product up to scale
Principal angles	Subspace pair	Inner product up to scale
Grassmann distances	Subspace pair	Inner product up to scale
Linear CKA	Representation pair	Normalized Gram geometry
Nearest-neighbour agreement	Representation pair	Distance ordering
Preserved under G_{iso}		
Norm, distance	Activations	Fixed Euclidean scale
Intervention magnitude	Displacement in V	Norm and scale
Absolute-threshold rank	Covariance spectrum	Absolute threshold
Procrustes distance	Representation pair	Euclidean alignment loss
Reconstruction loss	Residual in V	Euclidean residual norm
Preserved under signed permutations of the latent coordinates		
Coordinatewise sparsity penalty	Latent codes	Coordinatewise axes

Dictionary learning acts on two spaces. Under $h \mapsto Ah+t$, taking $D \mapsto AD$, $W_{\mathrm{enc}} \mapsto W_{\mathrm{enc}}A^{-1}$, and transforming the encoder and decoder biases to absorb t leaves every encoder preactivation, and hence the latent codes, unchanged. The reconstruction residual transforms as $h - \hat{h} \mapsto A(h - \hat{h})$, so the Euclidean loss is preserved for all residuals only when $A^\top A = I$. The restriction to G_{iso} on the activation side therefore comes from the reconstruction loss (Bricken et al., 2023; Templeton et al., 2024). On the latent side, $W \mapsto WB$ and $x \mapsto B^{-1}x$ leave the reconstruction unchanged, while a coordinatewise sparsity penalty reduces this mixing symmetry to signed permutations. These are distinct restrictions on distinct spaces.

6 Current Practice Leaves the Specification Implicit

The preceding sections make the specification explicit. We now examine what happens when its components remain implicit. The recurring problem is not that strong structural assumptions are necessarily unwarranted, but that objects, actions, and procedures are identified or changed without recording the corresponding change in the claim.

6.1 Identification by Storage Format

A linear probe produces a coefficient vector in V^*, while a steering method such as difference-in-means produces a displacement in V. Because both are stored as length-d arrays, they are routinely treated as the same kind of direction. Probe weights are compared to steering vectors by cosine, used as intervention directions, or combined with vectors of other provenance. For example, Bhalla et al. (2024) use both steering vectors and linear-probe weights as additive intervention directions and compare their directions by cosine similarity. The same identification appears across reading points. A direction estimated at one layer is often transported to another by the identity map because both residual streams have dimension d, even though they need not carry the same group action.

Proposition 3.2 separates two operations that this practice conflates. Once a metric is fixed, the induced map from V^* to V descends to projective spaces equivariantly under G_{sim}. A cosine comparison between a probe direction and a steering direction can therefore be meaningful under similarity geometry. Reusing the probe coefficients themselves as an additive displacement is stronger. The vector-level identification is equivariant only under G_{iso}.

6.2 Normalization and Calibration

Hidden group choices also enter through normalization and calibration. Contrastive Activation Addition estimates a difference-in-means vector and applies it as a scaled translation (Rimsky et al., 2024). The reported procedure normalizes vector magnitudes across behaviours but not across layers, where residual-stream norms grow over the forward pass. These choices assign meaning to length and therefore introduce Euclidean scale into an otherwise affine-covariant construction. Activation Addition adopts the opposite convention (Turner et al., 2023): its activation-difference vector is left unnormalized, so the displacement produced by a coefficient depends on the norm supplied by the sampled activations. The same numerical coefficient therefore does not denote the same displacement under the two conventions. Wollschläger et al. (2025) make the required calibration explicit by scaling an optimised refusal direction to match the norm of a difference-in-means direction, thereby stipulating a common Euclidean magnitude.

Because directional ablation depends only on the normalized direction, Wollschläger et al. (2025) sample unit directions within refusal cones directly rather than normalizing arbitrary convex combinations, which would bias the induced distribution over directions. In our terminology, the sampling procedure is adapted to the projective object actually consumed by the intervention.

6.3 Unnamed Reading Points

The reading point is itself part of the G-space. Corollary 4.2 shows that the same function-preserving parameter transformation can act trivially at the residual stream and as a nontrivial general linear transformation at a query or key site.

KeyDiff provides a direct case (Park et al., 2025a). The method evicts KV-cache entries using pairwise cosine similarity among keys within a head, and its supporting analyses report key and query cosines, their ℓ^2 norms, PCA of key caches, and $\log\det(KK^\top)$, all in per-head query or key space. The empirical performance of the eviction rule is not in question here. What is at issue is the status of the geometric quantities offered in explanation of it. In ordinary dot-product attention, the functionally relevant quantity is the bilinear pairing $q^\top k$, equivalently $W_Q^\top W_K$ at the parameter level (Elhage et al., 2021); with RoPE, it is the family $W_Q^\top R(\tau) W_K$ indexed by relative position. By Corollary 4.2 and Appendix C, this gauge is anisotropic even under RoPE, so none of these quantities is invariant. The quantities used to explain the computation are therefore not invariants of the computation they explain.

Related analyses of head-space quantities inherit the same dependence on chosen head coordinates. Cross-Gram singular values between head projections are not invariant under general invertible changes of basis, although the underlying residual-stream subspaces are, unless the spanning matrices are first orthonormalized (FokouĂŠ, 2026). Likewise, cosine comparisons of head-space singular vectors and their calibration against a fixed Euclidean reference distribution depend on the chosen head-space geometry (Ge et al., 2026). Appendix C.4 gives corresponding transformations.

By contrast, Yamagiwa et al. (2026) compare column spaces of transposed projection matrices, which are subspaces of the residual stream. Under $W_Q \mapsto \Lambda W_Q$, for example, $W_Q^\top \mapsto W_Q^\top \Lambda^\top$ leaves its column span unchanged. In our framework, the comparison targets an object preserved by the parameter symmetry rather than geometry internal to the gauge-dependent head coordinates.

Gauge-related parameter settings compute the same model function but can cause a cosine-based key-space eviction rule to select different tokens. Naming the reading point is therefore necessary to determine whether a reported geometric property belongs to the model, to an explicitly chosen geometry, or only to the sampled parameterization.

6.4 Inconsistency Within a Single Analysis

The clearest cases arise when the specification changes within one pipeline. Proposition 4.3 requires the conclusion to respect the actions introduced throughout the analysis, not merely the symmetry of its initial estimator. Three patterns recur.

One estimate, different G-spaces. Arditi et al. (2024) use one difference-in-means estimate both as an unnormalized displacement in V and as a projective direction in $\mathbb{P}(V)$ (Section 1). The estimate changes G-space between uses, and the combined conclusion is stronger than either use establishes.

Change of object mid-analysis. A probe coefficient belongs to V^*, while an additive intervention consumes an element of V, so moving from probing to intervention requires specifying how the two spaces are related. Li et al. (2023) compare probe-weight, mass-mean-shift, and CCS directions as alternatives for intervention after using probes to identify relevant attention heads. Their intervention partially makes the required calibration explicit: the chosen direction is normalized, and the added displacement is scaled by the empirical standard deviation of activations along that direction. Under a similarity $A = sQ$, a normalized probe direction transforms as $Q\hat{w}$ while the projected standard deviation transforms as $s\sigma$, so the calibrated displacement $\sigma\hat{w}$ transforms covariantly as $A(\sigma\hat{w})$. The complete construction is therefore compatible with G_{sim}, while identifying a probe direction in V^* with an intervention direction in V still depends on the chosen similarity geometry.

Preprocessing that restricts the group. SVCCA provides the failure at the level of procedures (Raghu et al., 2017). CCA alone is affine-invariant on centered representations, but SVCCA first performs singular-value truncation, a spectral operation not preserved by general invertible linear transformations. The CCA stage cannot restore the affine invariance lost in preprocessing.

In each case the problem is not the presence of a strong assumption. It is that the assumption enters at one stage while the conclusion is stated as though it inherited the weaker assumptions of another.

7 The Same Omission Beyond Linear Representation

The same specification issue appears in the Platonic Representation Hypothesis (Huh et al., 2024), although the hypothesis itself is unrelated to linear representation. Representations of models with different widths need not lie in a common representation space, so cross-model comparison proceeds through structures induced on shared inputs, such as neighbourhoods or similarity matrices. Because two representations may induce the same structure on one input distribution and differ elsewhere, the resulting claim is also indexed by the distribution on which the comparison is made. These dependencies are typically left implicit.

Convergence then amounts to saying that differently realized representations belong to a common equivalence class. The object, predicate, and comparison procedure are supplied by the hypothesis, while the transformation group remains open, and it is this remaining component on which the truth value turns. On a finite sample, sufficiently permissive linear equivalence can become degenerate. Kornblith et al. (2019) show that when the representation width is at least the number of sampled

inputs and both activation matrices have full row rank, a similarity measure invariant to arbitrary invertible linear transformations cannot distinguish them. Finite-sample convergence under such an equivalence can therefore become near-trivial in this regime. Under an isometric interpretation, by contrast, convergence would require agreement in absolute metric structure that models with different widths, tokenizers, and objectives generally cannot satisfy. Existing evaluations instead operate closer to G_{sim}, with the effective equivalence determined by the alignment measures in use rather than specified independently of them.

This places the Platonic Representation Hypothesis directly inside the audit of Section 5. Nearest-neighbour agreement, kernel-based alignment, and angular quantities inherit the same geometric commitments already catalogued there. The truth of the Platonic Representation Hypothesis cannot be assessed until the object being compared and the equivalence under which convergence is asserted have been specified.

8 The Specification Principle

One clarification is needed before stating the principle. Condition (A1) can always be satisfied by shrinking G or weakening P, so admissibility alone carries no content. What carries content is maximality: the strength of a claim is the largest group under which it survives. Reporting a predicate under an unnecessarily small group is not an error of validity but a loss of content. This is why Table 1 is organized by the largest preserving group rather than by any group under which a quantity happens to be defined.

We therefore propose a simple reporting principle. A representation study should state the G-space in which its object lives, the procedure F that produces the object and why it is equivariant under the declared action, and the predicate P ultimately asserted of it. It should also name the reading point at which the representation is taken and check the symmetry of the complete analysis when several stages are composed. These requirements need not be presented in group-theoretic notation, but the underlying choices must be recoverable from the statement of the claim rather than reconstructed from the implementation.

This requirement has a real cost. Some conclusions become narrower once their geometric assumptions are made explicit, some comparisons require an additional map between spaces, and some quantities must be attributed to a representation together with a chosen metric or parameterization rather than to the representation alone. One might instead declare G_{iso} throughout, but Condition (A2) forbids this when the architecture realizes transformations outside G_{iso}.

A second objection is that training returns one specific θ, so the geometry of that θ is a fact about the artifact one actually possesses, whatever the orbit contains. This is correct, and the framework does not forbid it. It fixes what such a statement is about. A property that varies across $\mathrm{Im}\,\rho$ is a property of θ together with the procedure that produced it, not of f_θ, and asserting it therefore incurs an empirical obligation that is rarely discharged: the reported geometry must be shown stable across seeds and training runs before it can be attributed to anything more general than the run in hand.

A third objection is that the proposal merely relabels assumptions already present in existing analyses. Park et al. (2024; 2025b), for example, make the geometry relating measurement and intervention explicit by specifying an inner product under which causally separable concepts are orthogonal. Golechha et al. (2025) argue that near-orthogonality can arise generically in high dimensions and be induced by whitening, so the resulting geometry is not uniquely diagnostic of learned conceptual structure. Making the geometry explicit therefore does more than relabel an assumption: it turns the disagreement into a precise question about which inner product supports the orthogonality predicate.

9 Conclusion

The Linear Representation Hypothesis is not one hypothesis but a family, and its members are distinguished by which representations they treat as the same. Stating a member requires a G-space, an equivariant procedure producing the object, and a predicate asserted of it, subject to the floor the architecture already fixes. Once these are recorded, two studies can be checked for whether they assert the same claim without reconstructing both analyses. Nothing in the requirement depends on linearity, and the same omission appears wherever a claim about representations is made at all.

AI USE STATEMENT

Generative AI tools were used to provide feedback on and refine the conceptual framework, and to edit the manuscript for clarity and readability. All AI-assisted suggestions were reviewed and selectively incorporated by the authors.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Usha Bhalla, Suraj Srinivas, Asma Ghandeharioun, and Himabindu Lakkaraju. Towards unifying interpretability and control: Evaluation via intervention. *arXiv preprint arXiv:2411.04430*, 2024.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Lucia Domenichelli, Dominique Brunato, and Felice Dell'Orletta. Linguistic profiling of transformer embedding geometry. In *Proceedings of the 30th Conference on Computational Natural Language Learning*, 2026. URL https://aclanthology.org/2026.conll-main.10/.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Josh Engels, Eric Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, volume 2025, pp. 84591–84622, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d3221cdb27e49d9c1cd35ad254feccfe-Paper-Conference.pdf.

Ernest FokouĂŠ. Multi-head attention as ensemble nadaraya-watson estimation: Variance reduction, decorrelation, and optimal head diversity. *arXiv preprint arXiv:2605.20271*, 2026.

Nikhil Garg, Jon Kleinberg, and Kenny Peng. How many features can a language model store under the linear representation hypothesis? *arXiv preprint arXiv:2602.11246*, 2026.

Yuyao Ge, Shenghua Liu, Yiwei Wang, Tianyu Liu, Baolong Bi, Lingrui Mei, Jiayu Yao, Jiafeng Guo, and Xueqi Cheng. Prism-Δ: Differential subspace steering for prompt highlighting in large language models. *arXiv preprint arXiv:2603.10705*, 2026.

Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11893–11905, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4df3510ad02a86d69dc32388d91606f8-Paper-Conference.pdf.

Satvik Golechha, Lucius Bushnaq, Euan Ong, Neeraj Kayal, and Nandi Schoots. Intricacies of feature geometry in large language models. In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL https://openreview.net/forum?id=Ut3ml7Hdwx.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Linfeng Liu, Tiffany Zhan, Louie Hong Yao, Saptarshi Ghosh, and Tianyu Jiang. Cross-lingual steering for figurative language generation. *arXiv preprint arXiv:2605.30443*, 2026.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, 2013. URL https://aclanthology.org/N13-1090/.

Chris Olah. What is a linear representation? what is a multidimensional feature? Transformer Circuits Thread, July 2024. URL https://transformer-circuits.pub/2024/july-update/index.html. Edited by Adam Jermyn.

Junyoung Park, Dalton Jones, Matthew J Morse, Raghavv Goel, Mingu Lee, and Christopher Lott. Keydiff: Key similarity-based KV cache eviction for long-context LLM inference in resource-constrained environments. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=uBaFH7aQnC.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 39643–39666, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 76441–76463, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/be7430d22a4dae8516894e32f2fcc6db-Paper-Conference.pdf.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6078–6087, 2017.

Shivam Raval, Hae Jin Song, Linlin Wu, Abir Harrasse, Jeff M Phillips, Fazl Barez, and Amirali Abdullah. Curveball steering: The right direction to steer isn't always linear. *arXiv preprint arXiv:2603.09313*, 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. URL https://aclanthology.org/2024.acl-long.828/.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=L9JM-pxQOl.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=80IwJqlXs8.

Hiroaki Yamagiwa, Yusuke Takase, and Hidetoshi Shimodaira. Measuring affinity between attention-head weight subspaces via the projection kernel. *arXiv preprint arXiv:2601.10266*, 2026.

Louie Hong Yao, Vishesh Anand, Yuan Zhuang, and Tianyu Jiang. Rhetorical questions in LLM representations: A linear probing study. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 155–172, 2026. URL https://aclanthology.org/2026.acl-long.5/.

Binchi Zhang, Zaiyi Zheng, Zhengzhang Chen, and Jundong Li. Beyond the permutation symmetry of transformers: The role of rotation for model fusion. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=wBJIO15pBV.

A Notation and Mathematical Preliminaries

A.1 Groups and Symmetries

A symmetry is a transformation that preserves a specified structure. Mathematically, a collection of compatible transformations is described by a group. The choice of group determines which transformations are regarded as equivalent descriptions of the same object.

A group G act on a set $\mathcal{X}$, if each element $g \in G$ assigns a transformation

$$x \mapsto g \cdot x \tag{5}$$

such that the identity transformation acts trivially and composition of transformations follows the group operation. The orbit of an element x under G is

$$[x]_G = \{g \cdot x : g \in G\}. \tag{6}$$

The orbit defines the equivalence relation induced by the group action: two elements are equivalent if they belong to the same orbit.

A.2 Linear Spaces

Let V be a finite-dimensional real vector space, and the dual space of V is the vector space of linear functionals

$$V^* = \{w : V \to \mathbb{R} \mid w \text{ is linear}\}. \tag{7}$$

A vector $v \in V$ and a dual vector $w \in V^*$ are related through the natural pairing

$$\langle w, v \rangle = w(v). \tag{8}$$

Although both V and V^* are finite-dimensional vector spaces with the same dimension, they are distinct spaces without additional structure such as an inner product.

An affine space is a vector space without a fixed origin, in which the difference between points forms vector, and the addition of a point and a vector results in a new point. Given a vector space V, an affine transformation takes the form

$$h \mapsto Ah + t, \tag{9}$$

where h and t are vectors in V, and A is an invertible linear transformation on V. Under an affine transformation, displacements transform as

$$(h_2 - h_1) \mapsto A(h_2 - h_1), \tag{10}$$

so affine relations are preserved while lengths and angles are generally not.

The projective space of V is the space of one-dimensional linear subspaces of V

$$\mathbb{P}(V) = \{\text{span}(v) : v \in V, v \neq 0\}. \tag{11}$$

Equivalently, projective space identifies vectors that differ only by a nonzero scalar multiplication. Therefore, quantities defined on $\mathbb{P}(V)$ describe directions rather than vector magnitudes.

The Grassmannian $\text{Gr}(k, V)$ denotes the space of all k-dimensional linear subspaces of V

$$\text{Gr}(k, V) = \{U \subseteq V : \dim(U) = k\}. \tag{12}$$

The projective space $\mathbb{P}(V)$ is the special case $\text{Gr}(1, V)$, where an element corresponds to a one-dimensional feature. More generally, a multidimensional linear feature is naturally represented as an element of $\text{Gr}(k, V)$.

For a vector space V, the general linear group is

$$\text{GL}(V) = \{A : V \to V \mid A \text{ is linear and invertible}\}. \tag{13}$$

It acts naturally on vectors by $v \mapsto Av$. The induced action on the dual space is $w \mapsto A^{-\top} w$ which preserves the vector-dual pairing.

An inner product on V is a positive-definite bilinear form that induces norms, distances, and angular quantities. For example,

$$\|v\| = \sqrt{\langle v, v \rangle}, \tag{14}$$

or

$$\cos(u, v) = \frac{\langle u, v \rangle}{\|u\| \|v\|}. \tag{15}$$

Such quantities depend on the chosen inner product and are therefore not intrinsic to a vector space alone.

B Proof of Proposition 3.2(i)

Let $f : V^* \to V$ satisfy $f(A^{-\top} w) = Af(w)$ for all $A \in \text{GL}(V)$ and all $w \in V^*$. Fix $w \neq 0$ and let $H_w = \{A \in \text{GL}(V) : A^{-\top} w = w\}$ be its stabilizer under the dual action, equivalently $\{A : w \circ A = w\}$. Equivariance gives $f(w) = Af(w)$ for every $A \in H_w$, so $f(w)$ lies in the subspace of V fixed pointwise by H_w.

Write $K = \ker w$, a hyperplane, and pick $u \notin K$. Every A preserving w acts arbitrarily on K and fixes u modulo K, so H_w contains all maps of the form $u \mapsto u + \kappa$, $A|_K \in \text{GL}(K)$ with $\kappa \in K$

arbitrary. Let $v \in V$ be fixed by all of H_w. Writing $v = \alpha u + \kappa_0$ with $\kappa_0 \in K$ and applying $u \mapsto u + \kappa$ gives $\alpha\kappa = 0$ for all $\kappa \in K$, hence $\alpha = 0$ whenever $\dim V \geq 2$. Then $v = \kappa_0 \in K$ is fixed by all of $\mathrm{GL}(K)$, which forces $\kappa_0 = 0$ since $\dim K \geq 1$ and $\mathrm{GL}(K)$ acts transitively on $K \setminus \{0\}$. Therefore $f(w) = 0$ for every $w \neq 0$, and $f(0) = 0$ follows from equivariance under $A = cI$ with $c \neq 1$. Hence $f \equiv 0$.

The hypothesis $\dim V \geq 2$ cannot be dropped. On $V = \mathbb{R}$ the dual action is $w \mapsto w/a$ and the map $f(w) = c/w$ satisfies $f(w/a) = af(w)$ for every $a \neq 0$, so nonzero equivariant maps exist in dimension one.

C Architecture-induced representation symmetries

C.1 The rotary commutant

Let $d_{\mathrm{head}} = 2n$ and let RoPE act at position t by $R(t) = \bigoplus_{j=1}^{n} R_2(t\theta_j)$, where $R_2(\alpha)$ is the planar rotation by α and $\theta_1, \ldots, \theta_n$ are the rotary frequencies. The attention logit between a query at position t and a key at position s is

$$\langle R(t)W_Q x, R(s)W_K y\rangle = x^\top W_Q^\top R(s-t)W_K\, y, \tag{16}$$

using $R(t)^\top R(s) = R(s-t)$.Writing $\tau = s-t$, a reparameterization $W_Q \mapsto \Lambda W_Q$, $W_K \mapsto \Lambda' W_K$ preserves every logit if and only if

$$\Lambda^\top R(\tau)\Lambda' = R(\tau) \qquad \text{for all relative positions } \tau. \tag{17}$$

Taking $\tau = 0$ gives $\Lambda' = \Lambda^{-\top}$, recovering the gauge of Corollary 4.2. Substituting back, Equation 17 becomes $R(\tau)\Lambda^{-\top} = \Lambda^{-\top}R(\tau)$ for all τ: the admissible $\Lambda^{-\top}$ are exactly the elements of the commutant of $\{R(\tau)\}_\tau$ in $\mathrm{GL}(2n)$.

Identify $\mathbb{R}^{2n} \cong \mathbb{C}^n$ by pairing the two coordinates of each rotary plane, under which $R(\tau)$ becomes multiplication by $\mathrm{diag}(e^{i\tau\theta_1}, \ldots, e^{i\tau\theta_n})$. When the frequencies are pairwise distinct and the relative positions τ range over enough values to separate them, the $\mathbb{R}$-algebra generated by $\{R(\tau)\}_\tau$ is the full diagonal algebra $\mathbb{C}^n$, whose commutant in $\mathrm{End}_{\mathbb{R}}(\mathbb{R}^{2n})$ is again $\mathbb{C}^n$. Its invertible elements are

$$\Lambda^{-\top} = \bigoplus_{j=1}^{n} s_j R_2(\alpha_j), \qquad s_j > 0,\ \alpha_j \in [0, 2\pi), \tag{18}$$

so $\mathrm{Im}\,\rho$ at a key site is the group of independently scaled rotations of the rotary planes, of real dimension $2n = d_{\mathrm{head}}$. Degenerate frequency sets enlarge the commutant: if m frequencies coincide, the corresponding block is $\mathrm{GL}_m(\mathbb{C})$ rather than $\mathbb{C}^\times$, so Equation 18 is the smallest image the architecture realizes.

C.2 Consequences at a key site

Equation 18 is a proper subgroup of $\mathrm{GL}(d_{\mathrm{head}})$, so the argument of Corollary 4.2 does not apply directly. It nonetheless excludes the metric predicates in question, because the scale factors s_j vary independently across planes. Write $k = (k^{(1)}, \ldots, k^{(n)})$ in rotary-plane blocks. Under Equation 18,

$$\|k\|^2 \mapsto \sum_j s_j^2 \|k^{(j)}\|^2, \tag{19}$$

which equals $\|k\|^2$ for all k only when every $s_j = 1$. Norms are therefore not invariant, and neither are cosines: choosing $s_1 \neq s_2$ and $s_j = 1$ otherwise changes the relative weight of the first two planes in $\langle k_1, k_2\rangle$ while changing the norms differently, so $\cos(k_1, k_2)$ is not preserved. The same anisotropy moves the eigenvalues of $KK^\top$ non-uniformly, so PCA directions, $\log\det(KK^\top)$, and any spectral quantity not expressible through ratios fixed by the per-plane scaling are likewise gauge dependent. Only a uniform scaling $s_1 = \cdots = s_n$ would restrict the action to G_{sim} and rescue the angular quantities, and the architecture does not impose it.

By contrast, the criterion of Section C.3 survives, since Equation 18 is a subgroup of the gauge under which Ω and the key differences were shown to transform oppositely. The RoPE restriction therefore shrinks the architectural image without restoring any of the quantities Corollary 4.2 excludes.

C.3 What an invariant version looks like

The framework does more than reject quantities. Because the QK gauge preserves realized attention logits $q^\top k$ exactly, any predicate expressible through those logits is admissible at these sites. Without RoPE, the corresponding parameter-level object is $W_Q^\top W_K$; with RoPE, it is the family $W_Q^\top R(\tau) W_K$ indexed by relative position. This is enough to restate the eviction criterion invariantly. What KeyDiff needs is a notion of when two realized keys are interchangeable in their effect on attention, and that effect is mediated through $q^\top(k_1 - k_2)$. Define the second moment of the realized queries by

$$\Omega = \mathbb{E}\left[qq^\top\right], \qquad d(k_1, k_2)^2 = (k_1 - k_2)^\top \, \Omega \, (k_1 - k_2). \tag{20}$$

In the no-RoPE case, $\Omega = W_Q \Sigma W_Q^\top$, where Σ is the second moment of the residual-stream inputs on the query side. Under the induced gauge $q \mapsto \Lambda q$ and $k \mapsto \Lambda^{-\top} k$, we have $\Omega \mapsto \Lambda \Omega \Lambda^\top$ and $k_1 - k_2 \mapsto \Lambda^{-\top}(k_1 - k_2)$, so the two transformations cancel and d is unchanged. The quantity it measures is also the right one: $d(k_1, k_2)^2$ is the mean squared change in attention logit incurred by substituting one key for the other. When $\Omega \propto I$ the criterion reduces to Euclidean distance in key space, which orders pairs as cosine does when key norms are equal, so the original rule is recovered as the special case in which the realized query distribution is isotropic in the gauge that happens to have been sampled. The distinction is not merely formal: gauge-related parameter settings compute the same model function while producing different eviction sets under the cosine rule and identical eviction sets under d.

C.4 Head-space quantities

The same architectural gauge constrains geometric comparisons between attention heads. FokouĂŠ (2026) compute cross-Gram matrices between key-projection matrices and interpret their singular values in terms of principal angles between head subspaces. If $G_{hh'}$ denotes such a cross-Gram matrix, independent invertible changes of basis within the two heads give

$$G_{hh'} \mapsto B_h^\top G_{hh'} B_{h'}. \tag{21}$$

Its singular values are therefore not invariant under general invertible B_h and $B_{h'}$ unless the spanning matrices have first been orthonormalized. This does not imply that the underlying subspaces are gauge dependent: when represented in residual-stream coordinates, those subspaces are unchanged by invertible changes of basis within the heads.

Ge et al. (2026) compare leading left singular vectors of different heads by cosine similarity and calibrate the observed statistic against uniformly sampled unit vectors in a fixed Euclidean $\mathbb{R}^{d_{\text{head}}}$. Both constructions depend on the chosen head-space geometry. An anisotropic change of head coordinates changes the cosine statistic, while the uniform distribution on the Euclidean unit sphere is itself not preserved by such a transformation.