Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Are Representation Measurements Well Defined? Parallel Analysis Under Function-Preserving Reparameterizations

Abdullah Karasan

akarasan@umbc.edu

University of Maryland, Baltimore County

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11700990/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

When Are Representation Measurements Well Defined? Parallel Analysis Under Function-Preserving Reparameterizations

Abdullah Karasan[1*]

[1*]University of Maryland, Baltimore County, Baltimore, 21250, MD, USA.
[2]Leveragai, Inc., USA.

Corresponding author(s). E-mail(s): akarasan@umbc.edu;

Abstract

Measurements on neural representations are often taken to describe properties of the model. However, hidden units can be transformed, with corresponding weight changes, so that the model's input–output function is unchanged while a representation measurement changes. This paper considers orthogonal basis changes in column permutation parallel analysis. Parallel analysis chooses a number of principal components by comparing observed eigenvalues with a reference distribution constructed by shuffling each column independently. The shuffling preserves each coordinate's empirical distribution, but these distributions depend on the basis. For centred data without coordinate-wise standardisation, different numbers of components can therefore be chosen even though the observed covariance spectrum is unchanged. In general, a reference distribution cannot simultaneously preserve every coordinate's empirical distribution, remove covariance between coordinates, and transform consistently under all orthogonal changes of basis. Experiments on stored representations from five language models across three retrieval domains examine component counts and context classifications in a retrieved-context task. A parallel analysis-selected context subspace is used to score answer representations. The task distinguishes contexts constructed with and without annotated evidence. Across 75 model–domain–transformation combinations, the median proportions of contexts with changed component counts and classifications were 0.79 and 0.27. Component-count changes also occurred without standardisation. In a centring-only control, counts changed in 1141 of 1200 contexts. Gram-matrix scores were numerically stable under orthogonal basis changes and performed similarly on held-out data. These

results show that measurements should be checked for invariance under function-preserving transformations before being interpreted as properties of the model's input–output function.

Keywords: Parallel analysis, neural representations, function-preserving reparameterization, invariance, principal component analysis, language models

1 Introduction

Neural representations can be measured to compare models, infer dimensionality and understand their behaviour. However, different internal representations can produce exactly the same outputs. For example, the hidden units can be permuted, with matching changes to the weights, to leave the predictions unchanged. Furthermore, several Transformer architectures allow for orthogonal transformations which combine hidden coordinates but maintain inner products and lengths. With the required parameter adjustments, these changes preserve the mathematical input–output function (Dinh et al. 2017; Godfrey et al. 2022; Ashkboos et al. 2024; Liu et al. 2025). However, depending on what is being measured, such a change can alter the measurement. Representational similarity analysis (RSA) and centered kernel alignment (CKA) are invariant to particular transformations of the coordinates (Kriegeskorte et al. 2008; Kornblith et al. 2019). On the other hand, analyses of individual coordinates can be affected by permutations, combinations and scaling of the units (Antverg and Belinkov 2022; Roeder et al. 2021). If the measurement is thought of as a property of the input–output mapping then it should be invariant to transformations that preserve this mapping.

In this paper we investigate whether the number of components selected by column permutation parallel analysis (PA) differs when basis transformations are applied to the hidden representations. Parallel analysis determines how many principal components to keep based on comparisons between the observed eigenvalues and a reference (Horn 1965; Buja and Eyuboglu 1992; Dobriban 2020). The representations are put into a matrix where each row represents an observation and each column represents a hidden coordinate. To build the reference, each column is randomly permuted independently, which keeps its values but removes associations with other columns. This means the reference preserves each coordinate's empirical marginal distribution exactly. Only components whose eigenvalue is greater than the corresponding threshold from the reference are kept.

When an orthogonal basis transformation is applied, if the data is centred and there is no coordinate-wise standardisation, then the singular values, overall variance and covariance eigenvalues are unchanged. However, the distribution of each column can change, and so the reference built from those columns can change too. This can mean that PA will select a different number of components, even though the inputs, outputs and mathematical mapping are kept constant. As a simple example (Theorem 2), take data that varies along a single coordinate. Shuffling that column does not change its variance, so the reference has the same largest eigenvalue as the data and PA keeps

no components. If the same data is rotated so that the variance is spread over three coordinates, then, with enough observations, the reference threshold for the largest eigenvalue is smaller than the observed eigenvalue and PA keeps one component.

Previous work has identified parameter symmetries, and has examined which transformations a representation comparison should be invariant to (Dinh et al. 2017; Godfrey et al. 2022; Ding et al. 2021; Hayne et al. 2024; Klabunde et al. 2025). This leads to two questions. First, can we find a reference that preserves coordinate marginals and removes dependence between coordinates, but is also equivariant to all orthogonal transformations? Equivariance here means that applying the transformation to the data before drawing the reference gives the same distribution as drawing the reference first and then applying the transformation. Second, can the choice of basis affect downstream scores and classifications, and not only the number of components?

We answer the first question theoretically. To test the second question, we look at retrieved context scoring. PA determines a subspace from the context token representations, and the answer representations are projected onto this subspace to give a score. A fixed threshold is applied to the score so the context can be classified. The goal of the task is to distinguish contexts constructed with and without annotated evidence. The number of components, score and classification are checked to see whether they change when applying a basis transformation. The inputs, answer, labels and mathematical mapping are kept constant. This provides a controlled experiment linking the spectral reference to a score and classification. It is not meant to be a general retrieval measure. The general result only restricts the distribution of the reference matrix; it does not by itself show that any particular statistic changes. Therefore, specific results and experiments are provided for PA, looking at the number of components, score and classification.

The paper makes three contributions.

1. **A general restriction on empirical references.** If a reference based only on the observed matrix preserves every coordinate marginal exactly and is right orthogonally equivariant, then it can only reorder whole rows. It therefore preserves the complete empirical row distribution, including the covariance between coordinates.
2. **Finite-sample results for PA.** Bounds on the permutation reference show that the number of components selected can change with the basis whilst the observed spectrum stays constant. Results are provided for the full orthogonal group used with RMSNorm models and for the LayerNorm compatible subgroup.
3. **An empirical test of counts, scores, and classifications.** Representations from five models in three retrieval domains are tested, with controls for the permutation seed, preprocessing and analysis settings. The results are compared to Gram matrix scores that are invariant to orthogonal basis changes. In the standardised analysis, both the observed spectrum and the reference can change, so the experiments do not show that the number of components is the sole cause of differing decisions.

2 Related Work

2.1 Representation Invariance and Measurement

Symmetries in neural networks cause different parameterisations to produce the same input-output function. For example, permutations of hidden units and compensatory scalings of the weights leave the function unchanged, which affects measures of the loss landscape and comparisons between trained networks (Dinh et al. 2017; Ainsworth et al. 2023; Godfrey et al. 2022). For instance, Dinh et al. (2017) show that scaling can be applied to maintain the function, but will change a sharpness metric. Techniques such as those in transformer compression use orthogonal transformations. QuaRot applies fixed rotations to neighbouring weights (Ashkboos et al. 2024) and SpinQuant learns the rotations, while keeping the full precision function intact (Liu et al. 2025). These transformations are used here to investigate whether the representation measure is dependent on the hidden basis. Representational Similarity Analysis (RSA) measures pairwise relationships between stimuli (Kriegeskorte et al. 2008). Linear centred kernel alignment (CKA) compares gram matrices across examples and is invariant to orthogonal transformations and isotropic scaling (Kornblith et al. 2019). Understanding of these metrics depends on what transformations they are sensitive to. Ding et al. (2021) test whether similarity measures are sensitive to changes that affect function and unaffected by changes that do not. Hayne et al. (2024) show that common similarity measures can provide a better measure of linearly decodable information compared to network information. ReSi investigates how 24 measures behave across language, vision and graph models using 6 different similarity criteria (Klabunde et al. 2025). Results from identifiability show that the model function cannot always be uniquely identified from the coordinates. Representations may be only identifiable up to an invertible linear transformation (Roeder et al. 2021) and representations may be far apart even if their outputs are almost identical (Nielsen et al. 2025). Interpretability studies have similar concerns. Probe accuracy can depend on the capacity of the probe (Hewitt and Liang 2019; Pimentel et al. 2020), individual neuron conclusions may depend on how the analysis is performed (Antverg and Belinkov 2022) and features in sparse autoencoders may not form a unique or complete set of units (Leask et al. 2025). This motivates making explicit assumptions about invariances. In this paper we look at a reference distribution from a single model. Our main result shows that a reference with exact marginal preservation and orthogonal equivariance must preserve the empirical row distribution. We then relate this to component choice and scoring using our results from PA.

2.2 Dimensionality Estimation in Language Models

Recent work has looked at understanding representations in language models by looking at their dimensionality. Cheng et al. (2025) found a high dimensional mid layer stage associated with linguistic abstraction, transferability and predictability across models. Lee et al. (2025) found non-linear intrinsic dimension correlates with semantic compositionality but linear dimensionality captures other input characteristics. Interpretation of these metrics requires clear understanding of what they are measuring.

Intrinsic dimension of a manifold, rank of covariance, effective rank and number of factors may differ. This paper looks at the number of factors chosen by permutation PA. Alternative methods include deterministic PA (Dobriban and Owen 2019) and random signflip PA (Hong et al. 2020). Hong et al. (2020) also show that, under heterogeneous noise, the permutation reference is governed by the average variance of each column rather than by the actual noise. Their concern is estimating the rank under heterogeneous noise. Ours is different: the coordinate variances themselves change under function-preserving changes of basis. We want to know if the number of factors as interpreted can be considered as a property of the model.

2.3 Context Sufficiency and Retrieval Evaluation

When evaluating retrieval tasks it is important to understand the quality of the evidence and how well the generator can utilise it. ARES measures context relevance, answer faithfulness and relevance (Saad-Falcon et al. 2024). RAGChecker provides diagnostics at the level of claims and is validated against human judgement (Ru et al. 2024). Joren et al. (2025) define sufficient context as having enough informative text to answer the query. This allows investigation of error and abstention. We call contexts constructed with annotated evidence evidence-containing and those constructed without it evidence-removal. These labels are assigned by construction and do not guarantee semantic sufficiency in the sense of Joren et al. (2025). This is a controlled experiment to investigate how the representation measures behave. We measure the performance of the score using held out AUROC to see how well we can separate the two conditions. We test if it depends on the hidden basis using reparameterisation.

3 Theoretical Framework

The theory outlines how inference from the representations takes place to produce the final score. First we define a function preserving change of basis. Then how many components PA picks out is investigated before and after the change of basis. After showing the general result about the constraints on reference distributions, we define scores that are independent of the hidden basis.

Let $X_\ell \in \mathbb{R}^{n_\ell \times d}$ be the representations of the retrieved context tokens at layer ℓ and $A_\ell \in \mathbb{R}^{m_\ell \times d}$ be the representations of the fixed target answer tokens. It is assumed that $n_\ell \geq 2$, $m_\ell \geq 1$ and there is some non-empty set of layers being scored. We apply the same change of hidden basis to both matrices:

$$(X_\ell, A_\ell) \longmapsto (X_\ell R, A_\ell R), \qquad \text{where } R \in O(d).$$

If we change the adjacent parameters of the model appropriately, this change of basis corresponds to a reparameterisation. The function of the model from input to output remains unchanged if there are no rounding errors. For the RMSNorm models described, the class of allowable transformations is the entire orthogonal group $\mathcal{G}_{\text{RMS}} = O(d)$, after making changes to account for the normalisation and weight sharing. The normalisation gains are absorbed into the adjacent weights, and the input embedding and output head are given separate parameters where necessary. This

does not change the function of the model but may have removed the constraint on the weight sharing. See Appendix B for details of the mappings. In the case of the LayerNorm model, the exact reparameterisations performed in our experiments are restricted to:

$$\mathcal{G}_{\text{LN}} = \{R \in O(d) : R\mathbf{1} = \mathbf{1}\}.$$

$\mathcal{G}_{\text{LN}}$ is the class of orthogonal transformations which preserve the mean direction, whilst also maintaining the function of the model. Theorem 3 requires the entire group $O(d)$, so it can be applied directly to the RMSNorm models but not to the LayerNorm model. For the LayerNorm model, Theorem 2 constructs an explicit example within $\mathcal{G}_{\text{LN}}$. We call these changes to the parameters function preserving reparameterisations, and we refer to the corresponding R as a change of basis or a transformation.

Once the class of transformations for each architecture has been defined, we say that a statistic $T(X, A)$ is invariant under the function preserving reparameterisations when,

$$T(XR, AR) = T(X, A) \qquad \text{for all } R \text{ in the class for the given architecture.}$$

If a decision is made based on this statistic with a fixed threshold τ, such that $D_\tau(X, A) = \mathbf{1}\{T(X, A) \geq \tau\}$, then it will also be invariant.

3.1 Coordinate Dependence of Column-Permutation Parallel Analysis

PA uses both the observed eigenvalues and a reference distribution. An orthogonal change of basis preserves the observed eigenvalues, but this alone does not preserve the selected count: the reference thresholds can change. The following results show how this happens.

PA is considered first because its reference covariance can be analyzed for both normalization architectures. After establishing the PA-specific result, Section 3.2 shows that, under the full orthogonal group, exact marginal preservation and equivariance also prevent more general reference procedures from changing the covariance.

To analyze the reference, let $X \in \mathbb{R}^{n\times d}$, $n \geq 2$, have centered columns and define

$$S = \frac{1}{n} X^\top X.$$

Here, $\lambda_k(A)$ denotes the k-th largest eigenvalue of a symmetric matrix A. For each feature j, independently draw a uniform $n \times n$ permutation matrix P_j and set $X^\pi_{\cdot j} = P_j X_{\cdot j}$. The nonparametric column-permutation form of parallel analysis is considered (Horn 1965; Buja and Eyuboglu 1992; Dobriban 2020). It compares the ordered eigenvalues of S with componentwise quantiles of

$$S^\pi = \frac{1}{n} (X^\pi)^\top X^\pi.$$

We call PA ideal when it uses the exact quantiles $t_k(X)$ of this permutation distribution, rather than estimates from a finite number of permutations. The estimator

retains consecutive components from the largest eigenvalue onward and stops at the first failed comparison:

$$\widehat{r}_{\mathrm{PA}}(X) = \max\bigl\{r \in \{0, \dots, \min(n,d)\} : \lambda_k(S) > t_k(X) \text{ for all } 1 \le k \le r\bigr\}.$$

Lemma 1 (Expected Permutation Covariance) *Conditionally on X,*

$$\mathbb{E}_\pi[S^\pi] = \mathrm{diag}(S).$$

For every $R \in O(d)$, X and XR have the same singular values, but

$$\mathbb{E}_\pi[S_R^\pi] = \mathrm{diag}(R^\top S R),$$

where S_R^π is generated by independently permuting the columns of XR. Thus the expected permutation-reference covariance can depend on the hidden-coordinate system even though the observed covariance eigenvalues do not.

Proof Write $x_j = X_{\cdot j}$. The diagonal entries satisfy

$$\frac{1}{n}(P_j x_j)^\top (P_j x_j) = \frac{1}{n}\|x_j\|_2^2 = S_{jj}.$$

For $j \neq k$, independence of P_j and P_k, together with column centering, gives

$$\mathbb{E}_\pi[(P_j x_j)^\top (P_k x_k)] = 0.$$

This proves the first identity. Applying it to XR proves the second, whereas $(XR)^\top (XR)/n = R^\top S R$ is orthogonally similar to S. □

Lemma 1 shows that the expected reference covariance depends on the covariance diagonal, which can change with the basis. However, PA uses quantiles of reference eigenvalues, not just their mean. The next lemma bounds deviations from the expected covariance so that the mean calculation can be used to bound those quantiles.

Lemma 2 (Permutation Covariance Bound) *Define*

$$V(S) = (\mathrm{tr}\, S)^2 - \sum_{j=1}^{d} S_{jj}^2.$$

Then, conditionally on X,

$$\mathbb{E}_\pi\Bigl[\bigl\|S^\pi - \mathrm{diag}(S)\bigr\|_{\mathrm{F}}^2\Bigr] = \frac{V(S)}{n-1}.$$

Consequently, for every $\delta \in (0,1)$, with conditional probability at least $1-\delta$,

$$\bigl\|S^\pi - \mathrm{diag}(S)\bigr\|_{\mathrm{op}} \le \varepsilon_\delta(S) := \sqrt{\frac{V(S)}{(n-1)\delta}}.$$

If $s_{(1)} \ge \cdots \ge s_{(d)}$ are the ordered diagonal entries of S, then simultaneously for all k,

$$\bigl|\lambda_k(S^\pi) - s_{(k)}\bigr| \le \varepsilon_\delta(S).$$

Define the upper $(1-\alpha)$-quantile by

$$Q_{1-\alpha}(Y) = \inf\{y : \mathbb{P}(Y \le y) \ge 1-\alpha\}.$$

Then the ideal componentwise threshold $t_k(X) = Q_{1-\alpha}\{\lambda_k(S^\pi)\}$ satisfies

$$t_k(X) \le s_{(k)} + \varepsilon_\alpha(S).$$

The proof of Lemma 2 is given in Appendix A.1. The $1/\sqrt{\delta}$ factor comes from Markov's inequality applied to the squared Frobenius norm. At $\delta = 0.05$ it is $\sqrt{20} \approx 4.47$ times the root-mean-square Frobenius fluctuation. The additional substitution $\|\cdot\|_{\mathrm{op}} \leq \|\cdot\|_{\mathrm{F}}$ can only make the bound looser. The bound is therefore a sufficient finite-sample guarantee, not an estimate of the actual 95th-percentile deviation.

The next step is to compare the thresholds in different bases. The Schur–Horn theorem describes the possible covariance diagonals. One basis aligns a coordinate with a leading eigenvector; another spreads the variance so that every diagonal entry is equal. These two choices give the following range for the largest eigenvalue of the expected reference covariance. If $S \succeq 0$, $\mathrm{tr}(S) > 0$, and $\bar{\lambda} = \mathrm{tr}(S)/d$, then

$$\min_{R \in O(d)} \lambda_1(\mathbb{E}_\pi[S_R^\pi]) = \bar{\lambda}, \qquad \max_{R \in O(d)} \lambda_1(\mathbb{E}_\pi[S_R^\pi]) = \lambda_1(S).$$

The lower endpoint is attained by a basis with constant covariance diagonal, and the upper endpoint is attained by aligning one coordinate with a leading eigenvector. Hence the largest eigenvalue of the expected reference covariance can vary over an exact multiplicative range of

$$\frac{d\,\lambda_1(S)}{\mathrm{tr}(S)}.$$

This range concerns the largest eigenvalue of the expected reference covariance, rather than the mean or a quantile of the largest eigenvalue across realized permutation references. This range for the expectation does not establish that PA thresholds or selected component counts must vary. Combining the Schur–Horn construction with the finite-sample quantile bound in Lemma 2 gives the spectrum-dependent threshold bound

$$c_\alpha(S, n) = \bar{\lambda} + \sqrt{\frac{(\mathrm{tr}\, S)^2 - (\mathrm{tr}\, S)^2/d}{(n-1)\alpha}}.$$

The number of leading eigenvalues exceeding this bound is

$$m_\alpha(S, n) = \max\left(\{0\} \cup \{m \in \{1, \ldots, \min(n, d)\} : \lambda_m(S) > c_\alpha(S, n)\}\right).$$

Here $m_\alpha(S, n)$ counts how many leading observed eigenvalues exceed the upper bound on the reference threshold in a basis with a constant covariance diagonal. In the eigenvector-aligned basis, the first PA comparison fails. Comparing these bases therefore gives a guaranteed difference in selected counts, as stated next.

Theorem 1 (Component-Count Range) *Consider ideal PA based on independent column permutations. Use upper $1-\alpha$ componentwise quantiles, selection from the largest component until the first failed comparison, and strict exceedance. Then*

$$\max_{R \in O(d)} \widehat{r}_{\mathrm{PA}}(XR) - \min_{R \in O(d)} \widehat{r}_{\mathrm{PA}}(XR) \geq m_\alpha(S, n).$$

In particular, a basis aligned with a leading eigenvector has selected component count zero, while a basis with a constant covariance diagonal retains at least $m_\alpha(S, n)$ components.

Appendix A.2 gives the proof. Because Lemma 2 is conservative, $m_\alpha(S, n)$ can be zero even when the actual PA count varies across orientations. Whenever it is positive, however, the theorem guarantees a positive range of counts across orthogonal bases, using only the observed spectrum. This result uses the full group $O(d)$. The next result provides an explicit construction that also lies within the LayerNorm-compatible subgroup.

Theorem 2 (Same Spectrum, Different PA Component Counts) *Fix integers $n \geq 2$ and $d \geq 3$, $\lambda > 0$, and $\alpha \in (0, 1)$. Let $u \in \mathbb{R}^n$ be any centered unit vector, e_1 the first coordinate vector, and*

$$w = (2/3,\ 2/3,\ -1/3,\ 0, \ldots, 0)^\top.$$

Define

$$X_{\mathrm{loc}} = \sqrt{n\lambda}\, u e_1^\top, \qquad X_{\mathrm{mix}} = \sqrt{n\lambda}\, u w^\top.$$

There exists $R \in \mathcal{G}_{\mathrm{LN}}$ such that $X_{\mathrm{mix}} = X_{\mathrm{loc}} R$, hence the matrices are related by a LayerNorm-compatible transformation and have the same singular values. Nevertheless, ideal PA with consecutive-from-largest selection and strict exceedance satisfies

$$\widehat{r}_{\mathrm{PA}}(X_{\mathrm{loc}}) = 0.$$

If

$$n - 1 > \frac{48}{25\alpha},$$

then

$$\widehat{r}_{\mathrm{PA}}(X_{\mathrm{mix}}) = 1.$$

The result is proved in Appendix A.3. Together, the two theorems provide complementary conclusions. Theorem 1 gives a lower bound on the range of counts over all orthogonal bases, whereas Theorem 2 gives a concrete count change within the LayerNorm-compatible subgroup. In both cases, the observed spectrum remains fixed while the permutation reference changes with the basis.

These results concern ideal PA with consecutive-from-largest selection and strict exceedance, a rule that may return zero. They establish coordinate dependence of the PA estimator rather than ambiguity in algebraic or covariance rank. Increasing the permutation budget estimates each coordinate-specific reference distribution more accurately but does not make those distributions equal.

The scorer used in the experiments additionally standardizes coordinates, enforces a minimum component count of one, and converts the selected subspace into a context–answer score. These steps matter because a change in the count does not, by itself, establish a change in the score or classification. The full scoring procedure is therefore defined next, and the experiments test its outputs.

Specifically, the implemented PA projection scorer first maps

$$X \longmapsto \mathcal{S}(X) = (X - \mathbf{1}\bar{x}^\top) D_X^{-1},$$

where D_X contains coordinatewise sample standard deviations, each floored at 10^{-6}. In general,

$$\mathcal{S}(XR) \neq \mathcal{S}(X)R.$$

Consequently, under reparameterization, both the standardized observed spectrum and its permutation reference may change. The empirical study therefore tests stability of the full scoring procedure, not only the unstandardized covariance calculation in Lemma 1.

To connect component selection to the empirical task, define the complete PA projection score as follows. At layer ℓ, let

$$\widetilde{X}_\ell = (X_\ell - \mathbf{1}\bar{x}_\ell^\top)D_{X_\ell}^{-1}, \qquad \widetilde{A}_\ell = (A_\ell - \mathbf{1}\bar{x}_\ell^\top)D_{X_\ell}^{-1}.$$

Let $r_\ell^{\mathrm{PA}} \geq 1$ be the PA-selected component count after the implemented minimum-count rule, and let $V_\ell^{\mathrm{PA}} \in \mathbb{R}^{d \times r_\ell^{\mathrm{PA}}}$ contain the corresponding leading right singular vectors of $\widetilde{X}_\ell$. For target-answer token i, define the projected squared-norm fraction

$$q_{\ell i}^{\mathrm{PA}} = \frac{\|\widetilde{a}_{\ell i} V_\ell^{\mathrm{PA}}\|_2^2}{\|\widetilde{a}_{\ell i}\|_2^2}.$$

This fraction measures how much of an answer token's squared norm lies in the selected context subspace. The implementation floors the denominator at 10^{-12}, so a zero answer vector contributes zero. Writing Agg_i for the mean of the largest $\max\{1, \lceil m_\ell/4 \rceil\}$ token-level values, the layer score is

$$e_\ell^{\mathrm{PA}} = \mathrm{clip}_{[0,1]}\left(\frac{\mathrm{Agg}_i(q_{\ell i}^{\mathrm{PA}}) - r_\ell^{\mathrm{PA}}/d}{1 - r_\ell^{\mathrm{PA}}/d}\right).$$

The term r_ℓ^{PA}/d is the fraction expected for a random subspace of the same dimension (Section 3.3). If $r_\ell^{\mathrm{PA}} = d$, e_ℓ^{PA} is defined as one. The reported PA projection score is the geometric mean of the clipped layer scores,

$$s_{\mathrm{PA}} = \exp\left\{\frac{1}{|\mathcal{L}|}\sum_{\ell \in \mathcal{L}} \log\big(\mathrm{clip}_{[10^{-6},1]}(e_\ell^{\mathrm{PA}})\big)\right\}.$$

The selected component count and the projection score are different quantities. The experiments examine whether each changes under reparameterization. A change of basis can affect standardization, component selection, projection scoring, and the final classification.

3.2 Limits of Marginal-Preserving Reference Distributions

The preceding results concern the particular reference produced by column permutations. A natural question is whether another reference can keep the same coordinate marginals and remove cross-coordinate covariance without introducing dependence on the basis. This section considers reference procedures based only on the observed representation matrix and states precisely what consistency under a change of basis requires.

For a matrix X with rows x_i, define its empirical row measure and projected empirical measures by

$$\widehat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}, \qquad \widehat{\mu}_{Xv} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i^\top v}.$$

Let $\mathcal{K}(X, \cdot)$ denote a randomized reference procedure that uses only X to produce an $n \times d$ reference matrix, and write $Z \sim \mathcal{K}(X, \cdot)$. Restricting the procedure to the observed data simplifies the notation, but the proof does not require this restriction. For any fixed external input E, the same conclusion holds for $\mathcal{K}(X, E, \cdot)$ if exact marginal preservation holds for every X and right-orthogonal equivariance holds with E fixed. Thus, fixed external information does not change the conclusion if both requirements still hold.

The procedure preserves coordinate marginals exactly if, for every fixed X,

$$\widehat{\mu}_{Ze_j} = \widehat{\mu}_{Xe_j} \quad \text{almost surely for } j = 1, \ldots, d.$$

It is right-orthogonally equivariant if, for every fixed X and $R \in O(d)$,

$$\mathcal{K}(XR, \cdot) = \mathcal{L}(ZR \mid Z \sim \mathcal{K}(X, \cdot)),$$

where $\mathcal{L}$ denotes the distribution of a random matrix.

In words, changing the input basis before drawing a reference must give the same distribution as drawing the reference first and then changing its basis.

Although the assumptions concern coordinate marginals and changes of basis, together they determine the entire empirical row distribution.

Theorem 3 (Marginal-Preservation Limit) *Suppose a randomized reference procedure based only on the observed matrix preserves every coordinate marginal. Suppose also that, for every input matrix and every $R \in O(d)$, its output distribution transforms from Z to ZR. Then*

$$\widehat{\mu}_Z = \widehat{\mu}_X \quad \textit{almost surely.}$$

Equivalently, $Z = PX$ for a random row-permutation matrix $P = P(X, Z)$. Therefore,

$$Z^\top Z = X^\top X$$

almost surely.

The proof is given in Appendix A.4. The conclusion is concrete: the reference can reorder the observed rows, but it cannot change which coordinate values occur together in a row. This differs from shuffling columns independently, which breaks those pairings. The reference therefore preserves the complete empirical row distribution, including its covariance and dependence structure.

Corollary 1 (Limitation for Empirical References) *Any empirical reference procedure based only on the observed data that exactly preserves every coordinatewise empirical marginal distribution and is right-orthogonally equivariant must produce a reference matrix Z satisfying*

$$Z^\top Z = X^\top X \qquad \textit{almost surely.}$$

Since marginal preservation also preserves coordinate means, the centered covariance is unchanged. Such a procedure therefore cannot remove nonzero cross-coordinate covariances.

The stated theorem and corollary pertain to the distribution of the full reference matrix and so the equivariance is stronger than invariance for a scalar statistic. This means that a lack of reference equivariance does not guarantee that any particular statistic involving the reference will change. The theorems specific to PA show the chosen number of components can change. This corollary can be applied even if we are not using retrieved context. It can be applied anytime where independent coordinate shuffles are used to construct the null spectrum and then the resulting number of components, eigenspace, threshold or hypothesis is taken to be a property of the learned representation. This is not saying that column permutation PA is not useful for describing the data. However, exact marginal preservation is not sufficient to allow interpretation of the results as being basis independent, since the marginals change with function preserving basis changes.

Table 1 shows which properties hold for each of three reference methods. Column permuting produces a reference with changed covariance at the cost of orthogonal equivariance. A trace matched isotropic Gaussian reference can be equivariant with changed covariance at the expense of exact empirical marginal preservation. Common row permuting fulfils both requirements, and therefore leaves the covariance unchanged.

Table 1 Properties of reference procedures based only on the observed matrix. "Changed second moment" asks whether the procedure can produce $Z^\top Z \neq X^\top X$.

Reference procedure	Exact marginals	$O(d)$-equivariant	Changed second moment
Independent column permutations	Yes	No	Yes
Common row permutation	Yes	Yes	No
Trace-matched isotropic Gaussian	No	Yes	Yes

There are two points to note here. Firstly, exact preservation of coordinate marginals and right-orthogonal equivariance are both required. If we drop either condition then we can have a reference with changed covariance. Secondly, the proof relies on having the whole group $O(d)$, with rotations able to take a fixed coordinate axis and rotate it to any point on the unit sphere. This means that the result applies directly to the RMSNorm reparameterisations we study here. We cannot extend the result to $\mathcal{G}_{\mathrm{LN}}$ because it fixes the all ones direction.

The two normalisation cases will require different results. The general theorem applies to the whole orthogonal group used by our RMSNorm models. However, Theorem 2 provides a direct result for the LayerNorm compatible subgroup showing PA instability. In the next section, we define scores invariant under orthogonal basis changes.

3.3 Invariant Context–Answer Scores

The PA score depends on a reference constructed from individual coordinates. The comparison scores instead use inner products, which are preserved by orthogonal changes of basis. These inner products form row Gram matrices, as in representational similarity analysis and centered kernel alignment (Kriegeskorte et al. 2008; Kornblith et al. 2019). In particular, for $X, Y \in \mathbb{R}^{n\times d}$,

$$XX^\top = YY^\top \quad \Longleftrightarrow \quad Y = XR$$

for some $R \in O(d)$. Thus $XX^\top$ characterizes X up to a right-orthogonal transformation. Here, the same construction is used to compare context and answer representations within one model. The scores differ in how they select or weight context directions, but both use quantities that are unchanged by a common orthogonal change of basis.

For the paired object (X, A), applying the same fact to the vertically concatenated matrix gives the block Gram matrix

$$\begin{bmatrix} X \\ A \end{bmatrix} \begin{bmatrix} X \\ A \end{bmatrix}^\top = \begin{bmatrix} XX^\top & XA^\top \\ AX^\top & AA^\top \end{bmatrix}.$$

The comparison scores use $XX^\top$, $AX^\top$, and answer norms, all of which are blocks or diagonal entries of this matrix. These quantities are unchanged by a common orthogonal change of basis. Their ability to distinguish the context conditions is evaluated separately.

At each layer ℓ, both representation sets are centered using the context mean.

$$X_\ell^c = X_\ell - \mathbf{1}_{n_\ell}\bar{x}_\ell^\top, \qquad A_\ell^c = A_\ell - \mathbf{1}_{m_\ell}\bar{x}_\ell^\top.$$

Let

$$G_\ell = \frac{1}{n_\ell} X_\ell^c (X_\ell^c)^\top, \qquad G_\ell = \sum_{j=1}^{n_\ell} \gamma_{\ell j} u_{\ell j} u_{\ell j}^\top,$$

where $\gamma_{\ell 1} \geq \cdots \geq \gamma_{\ell n_\ell} \geq 0$ are the full Gram-matrix eigenvalues and the $u_{\ell j}$ are orthonormal. Define

$$\epsilon_{\mathrm{eig},\ell} = \epsilon_{\mathrm{mach}} \max\{n_\ell, d\} \max\{1, \gamma_{\ell 1}\}, \qquad p_\ell = \#\{j : \gamma_{\ell j} > \epsilon_{\mathrm{eig},\ell}\}.$$

Let $U_\ell = [u_{\ell 1}, \ldots, u_{\ell p_\ell}]$, $\lambda_{\ell j} = \gamma_{\ell j}$ for $j \leq p_\ell$, and $\Lambda_\ell = \mathrm{diag}(\lambda_{\ell 1}, \ldots, \lambda_{\ell p_\ell})$. Only these eigenpairs are retained. Their truncated product need not equal G_ℓ when discarded eigenvalues are nonzero.

If $p_\ell = 0$, both comparison layer scores are defined as zero. Otherwise, define

$$Z_\ell = A_\ell^c (X_\ell^c)^\top U_\ell (n_\ell \Lambda_\ell)^{-1/2}.$$

This is $A_\ell^c V_\ell$, where V_ℓ contains the corresponding right singular vectors of X_ℓ^c, written using only the row Gram matrix and context–answer inner products.

Let $r_{\ell,\eta}$ be the smallest r such that

$$\frac{\sum_{j=1}^{r} \lambda_{\ell j}}{\sum_{j=1}^{p_\ell} \lambda_{\ell j}} \geq \eta.$$

The variance-threshold score uses the smallest principal subspace that accounts for the specified fraction of retained variance. The invariance result below assumes a cutoff that does not split a repeated eigenspace:

$$\lambda_{\ell, r_{\ell,\eta}} > \lambda_{\ell, r_{\ell,\eta}+1},$$

where $\lambda_{\ell, p_\ell+1} = 0$. This condition makes the selected spectral projector uniquely defined. The numerical implementation uses the ordered eigendecomposition at the specified tolerance. For answer row $a_{\ell i}^c$, define

$$h_{\ell i} = \frac{\sum_{j=1}^{r_{\ell,\eta}} Z_{\ell,ij}^2}{\|a_{\ell i}^c\|_2^2}, \qquad g_{\ell i} = \frac{\sum_{j=1}^{p_\ell} \frac{\lambda_{\ell j}}{\lambda_{\ell j}+\rho_\ell} Z_{\ell,ij}^2}{\|a_{\ell i}^c\|_2^2}.$$

As implemented, each denominator is floored at 10^{-12}, so a zero centered-answer row contributes zero rather than an undefined ratio. Here $\eta = 0.9$, $\rho_\ell = 0.1 \max\{p_\ell^{-1} \sum_{j=1}^{p_\ell} \lambda_{\ell j}, 10^{-12}\}$, and Agg denotes the mean of the largest $\max\{1, \lceil m_\ell/4 \rceil\}$ token-level values. The layer scores are

$$e_\ell = \mathrm{clip}_{[0,1]} \left(\frac{\mathrm{Agg}_i(h_{\ell i}) - r_{\ell,\eta}/d}{1 - r_{\ell,\eta}/d} \right), \qquad g_\ell^\star = \mathrm{clip}_{[0,1]}(\mathrm{Agg}_i(g_{\ell i})).$$

Thus $h_{\ell i}$ uses the selected principal subspace, whereas $g_{\ell i}$ weights every retained context direction by its ridge factor. For a fixed nonzero answer vector and an unfloored norm denominator, the quantity $r_{\ell,\eta}/d$ is the expected projected fraction for a Haar-uniform random $r_{\ell,\eta}$-dimensional subspace of $\mathbb{R}^d$. It is not estimated from permuted data. The definition assumes $r_{\ell,\eta} < d$. If the selected subspace spans all d coordinates, define the variance-threshold score as one and omit the baseline correction. For a fixed set of measured layers $\mathcal{L}$, the reported scores are

$$s_{\text{variance}} = \exp \left\{ \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \log\big(\mathrm{clip}_{[\epsilon,1]}(e_\ell)\big) \right\},$$

$$s_{\text{ridge}} = \exp\left\{ \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \log\big(\text{clip}_{[\epsilon,1]}(g_\ell^\star)\big) \right\}, \qquad \epsilon = 10^{-6}.$$

By construction, the resulting scores are jointly right-orthogonally invariant. Specifically, for every $R \in O(d)$, every nonempty finite collection of layers, and every valid input pair, with the tie-safe cutoff condition imposed for s_{variance},

$$s_{\text{variance}}(\{X_\ell R, A_\ell R\}_{\ell \in \mathcal{L}}) = s_{\text{variance}}(\{X_\ell, A_\ell\}_{\ell \in \mathcal{L}}),$$

and

$$s_{\text{ridge}}(\{X_\ell R, A_\ell R\}_{\ell \in \mathcal{L}}) = s_{\text{ridge}}(\{X_\ell, A_\ell\}_{\ell \in \mathcal{L}}).$$

This property follows directly from

$$(X_\ell^c R)(X_\ell^c R)^\top = X_\ell^c (X_\ell^c)^\top, \qquad (A_\ell^c R)(X_\ell^c R)^\top = A_\ell^c (X_\ell^c)^\top,$$

together with preservation of centered answer norms. The baseline correction, clipping, token aggregation, and cross-layer geometric mean are deterministic, so they preserve the equality.

Repeated eigenvalues require care. We cannot get an invariant spectral projector purely using a numeric decomposition. The ridge score is invariant even with repeated eigenvalues, and the variance threshold score is invariant under the cutoff condition above. However, if the cutoff splits a repeated eigenspace, we cannot get a basis invariant mathematical definition using an ordered numeric decomposition.

4 Empirical Analysis

We perform experiments to see how the PA measurements depend on the choice of basis. Firstly, we compare the PA counts, scores and classifications before and after a basis change. We then run controls to check whether the changes can be explained by the random permutation seed or by coordinate-wise standardisation. Finally, we check that the scores are still able to differentiate between the context conditions and that the PA result holds under other PA settings.

4.1 Experimental Setup

The five models used in the study were chosen according to the recorded protocol: Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen3-0.6B, SmolLM2-1.7B and Pythia-1B (Qwen Team 2024, 2025; Ben Allal et al. 2025; Biderman et al. 2023). The transformation class for each model can be seen in Table 2. We performed the basis changes directly on the stored residual states. The parameter changes are checked separately in Appendix C.

We use three retrieval datasets, which we refer to as domains: FinanceBench, QASPER and HotpotQA (Islam et al. 2023; Dasigi et al. 2021; Yang et al. 2018). In FinanceBench, we match evidence-containing with evidence-removal contexts by taking evidence from the target filing and pairing it with an evidence-free passage from the same company but a different time. For QASPER, we pair evidence-containing contexts which have annotated evidence paragraphs with evidence-removal contexts

Table 2 Models and allowed changes of basis after normalization scales are absorbed into adjacent weights and, where needed, the output head is given separate parameters. Parameter counts are those encoded in the public model identifiers.

Model	Architecture	Normalization	Transformation class
Qwen2.5-0.5B	Qwen2	RMSNorm	$O(d)$
Qwen2.5-1.5B	Qwen2	RMSNorm	$O(d)$
Qwen3-0.6B	Qwen3	RMSNorm	$O(d)$
SmolLM2-1.7B	Llama-style	RMSNorm	$O(d)$
Pythia-1B	GPT-NeoX	LayerNorm	$\{R \in O(d) : R\mathbf{1} = \mathbf{1}\}$

which only have sampled non-evidence paragraphs (up to 3). For HotpotQA, we match evidence-containing and evidence-removal contexts.

The evidence-containing and evidence-removal contexts are created by adding annotated evidence to a common evidence-removal context. After this, we truncate each context to 192 tokens. The question and target answer are identical for the two contexts in a pair. Contexts are labelled with how they were created. However, some evidence may be cut out during truncation, and the evidence-removal context could contain different evidence-removal words to the evidence-containing context. For each model-domain pair, we sample 20 questions for calibration and 50 disjoint questions for evaluation. Each question gives one evidence-containing and one evidence-removal context, so there are 40 calibration contexts and 100 evaluation contexts. We apply teacher forcing to the same target words in both contexts. At 1/4, 1/2, and 3/4 points along the model depth, we extract the hidden states. The layer scores are then combined with a geometric mean.

As a behaviour baseline, we use the average log probability of the answer. This is calculated by averaging the teacher-forced log probabilities of the target words. A higher value indicates an evidence-containing context. The same calibration and evaluation process is performed as with the representation scores. To test the measurements under function-preserving changes of basis, we sample 5 different transformations for each model with fixed random seeds. These transformations are applied to all 3 domains for each model. In the case of RMSNorm with width $d = mb$, where b is the largest power of 2 which divides d, the transformation matrix has the form

$$R = (Q_m \otimes H_b)\Pi D.$$

Here H_b is a normalised $b \times b$ Hadamard matrix. Q_m is obtained through QR decomposition of an $m \times m$ standard Gaussian matrix, with the signs on the diagonal of the triangular matrix incorporated into Q_m. Π is a uniform random permutation of the coordinates. D is a diagonal matrix which contains independent Rademacher entries. This means that the RMSNorm transformation is drawn from the structured family above, instead of being Haar uniform in $O(d)$. In the case of the LayerNorm model, $Q_\perp$ is Haar uniformly sampled from $O(d-1)$. This is done through QR decomposition of a standard Gaussian matrix of size $(d-1) \times (d-1)$. The matrix $\mathrm{diag}(1, Q_\perp)$ is then

applied in an orthonormal basis whose first vector is $\mathbf{1}/\sqrt{d}$. The resulting matrix will fix $\mathbf{1}$, and will be Haar uniform in its orthogonal complement.

This gives 25 model–transformation pairs and 75 model–domain–transformation combinations, which are not independent because they share models, domains and transformations. Table 3 lists the units of analysis used in the rest of this section.

Table 3 Units of analysis. Each question gives two contexts, and each evaluation context is compared with its transformed version under each of the five transformations.

Quantity	Number
Models	5 (4 RMSNorm, 1 LayerNorm)
Domains	3
Model–domain pairs	15 (12 for the RMSNorm models)
Transformations per model	5
Model–domain–transformation combinations	75
Questions per model–domain pair	20 calibration, 50 evaluation
Contexts per model–domain pair	40 calibration, 100 evaluation
Original evaluation contexts	1,500 (1,200 for the RMSNorm models)
Matched original–transformed context comparisons	7,500

For the main analysis, hidden states are extracted once for each original model. These are then transformed to $X_\ell R$ and $A_\ell R$. The teacher-forced answer log probability baseline is computed once from the original model. It is not recomputed from the transformed checkpoints. The layer indices are $\lfloor L/4 \rfloor$, $\lfloor L/2 \rfloor$ and $\lfloor 3L/4 \rfloor$, where L is the depth of the model. These hidden states are residual states before the final normalization. The transformed states correspond to the analytical parameter mappings described in Appendix B.

Transformed checkpoints are tested on 10 fixed teacher forced prompts. We report the results of these checks in Appendix C. The main analysis therefore concerns the mathematical change of basis applied to the stored states. We do not confirm that all 25 transformed checkpoints meet the numeric requirements: 22 were checked and 21 passed.

PA uses 50 column permutation replicates for each transformed or original representation. For each representation, we look at

$$k = \min(64,\ \min(n_\ell, d) - 1)$$

components. The threshold for each component is the 0.95 empirical quantile of its reference eigenvalues, computed through linear interpolation as in `torch.quantile`. We select consecutive components from the largest with strict exceedance, up to at most k. If all k components pass, all k components are chosen. For the scoring method, we ensure that we select at least 1 component. The component count reported for a context is the median of the three per-layer counts. Once the scores have been calculated, we choose an individual threshold for each score. This is done by looking at the 40 original contexts in calibration. Candidate thresholds are the midpoint between each pair of consecutive unique scores, and 1 value below the lowest score, and 1 value

above the highest score. These are then tested for accuracy. If multiple thresholds give the highest accuracy, the lower median of these thresholds is chosen. This ensures that we are choosing the threshold that gives the highest accuracy in calibration. The threshold is selected on the original representations and is kept fixed when the representations are transformed. AUROC does not require any thresholding, and is calculated using the 100 contexts in evaluation for each model-domain pair.

Each outcome compares a context's original representation with its transformed representation. Our main outcomes are disagreement in the number of components chosen (the component count), the magnitude of the score difference, disagreement in the decision (the context classification) when using a fixed threshold, and a reversal in the order of the 2 contexts for the question. Reversal in the ordering means that the original and transformed scores disagree on whether the evidence-containing score is strictly greater than the evidence-removal score. Ties are given an indicator of 0. There are no tied pairs within the main evaluation scores.

For each question, we take the average of the score differences between the 2 matched contexts. As each question gives us a pair of matched contexts, we use the bootstrap to sample whole questions. If we sampled the contexts separately, we would assume independence between the two contexts of a pair. For the main stability analysis, we condition on the 5 fixed transformations. We sample questions as our independent unit. This means that we can see how much variation there is between questions. For the paired sampling, we keep the matched evidence-containing and evidence-removal contexts. For the controls and AUROC, bootstrap intervals are calculated by sampling from the model-domain pairs we observed (12 for the controls and 15 for AUROC). This means we know how much variation there is in the model-domain pairs we have looked at. As the model-domain pairs share models and domains, this does not reflect uncertainty due to sampling from the population. For the robustness analysis, we sample 2000 replicates at the question level. We keep the matched pairs of contexts. We share the indices sampled for the questions between models in the same domain. As the same questions are used for evaluation for each model, the number of contexts for each model does not represent independent linguistic examples.

4.2 Sensitivity to Changes of Basis

Table 4 indicates that the counts can often change, but typically only by a small number of components. From the 75 comparisons, we find that the median fraction of contexts with changed counts is 0.79. The median absolute change in counts across the 7,500 context matches is 2. The median count is 39 for the original and 37 for the transformed representations. This shows that the counts are sensitive to the sampled bases. The median disagreement rate for fixed threshold is 0.27 and the median reversal rate for within question ordering is 0.10. This shows that ordering is a more stable comparison than fixed threshold. We cannot infer that changes in counts are the cause of changes in classifications, because in the standardised analysis both the observed spectrum and the reference can change (Section 4.3).

Overall, the counts more often decreased. Of the 7,500 context matches, 5,025 had reduced counts, 1,238 increased and 1,237 remained the same. This describes the sampled transformations and does not show that PA counts will generally decrease under

reparameterisation. Theorem 2 tells us that there can be different counts, although does not predict whether they will increase or decrease for these models. Figure 1 shows how the rates vary between model-domain pairs, and Figure 2 shows how the counts and decisions have changed.

For the 1,200 original contexts from the four RMSNorm models, the bound from Theorem 1 was calculated for $\alpha = 0.05$. The bound was positive for 182 out of the 3,600 context-layer spectra (5.1%). These were from 182 different contexts (15.2% of the 1,200 contexts). This means that for these layers, the ideal centring-only PA will choose a differing number of counts for at least two orthogonal bases. For the remaining spectra the bound was zero, which is inconclusive: it does not show that their counts are invariant. A positive bound also does not guarantee that there will be a change in counts for any sampled transformation, or a change in the reported count, because the reported count is at least 1 and is the median across the layers. The bound was not applied to Pythia as the LayerNorm compatible transformations make up a proper subgroup of $O(d)$.

Table 4 Changes induced by the orthogonal transformations. The first three rows summarise rates across the 75 model–domain–transformation combinations. The final row is a summary of the 7,500 matched context comparisons. The IQR and range are descriptive statistics for the observed comparisons, they are not confidence intervals.

Outcome	Median	IQR	Range
PA component-count disagreement	0.79	[0.725, 1.000]	[0.59, 1.00]
Decision disagreement	0.27	[0.160, 0.345]	[0.00, 0.44]
Within-question ordering reversal	0.10	[0.060, 0.140]	[0.02, 0.24]
Absolute component-count change	2	[1, 4]	[0, 12]

The maximum absolute change in variance threshold and ridge scores was 5.41×10^{-8} and 4.68×10^{-8} respectively. This is within numerical error. The PA projection score has higher change than both comparison scores for all 50 questions across all 75 comparisons. Therefore these comparison scores are a stable baseline to compare PA sensitivity to. This control tests consistency of the measurement with a constant model function.

4.3 Control Analyses

The controls were run on 12 of the 15 model–domain combinations. The first control compares the effect of the hidden basis with the effect of the permutation seed. PA was run once more with no changes to activations, using a different seed and the same budget of 50 reference permutations. The seed only disagreement rate is then compared to the median disagreement rate over the five fixed basis changes for each combination. Table 5 displays the median disagreement rates of 0.06 for the seed only run, and 0.76 in reparameterisation. The median difference between the paired runs is 0.69. A 95% bootstrap interval using the model-domain bootstrap method gives [0.625,

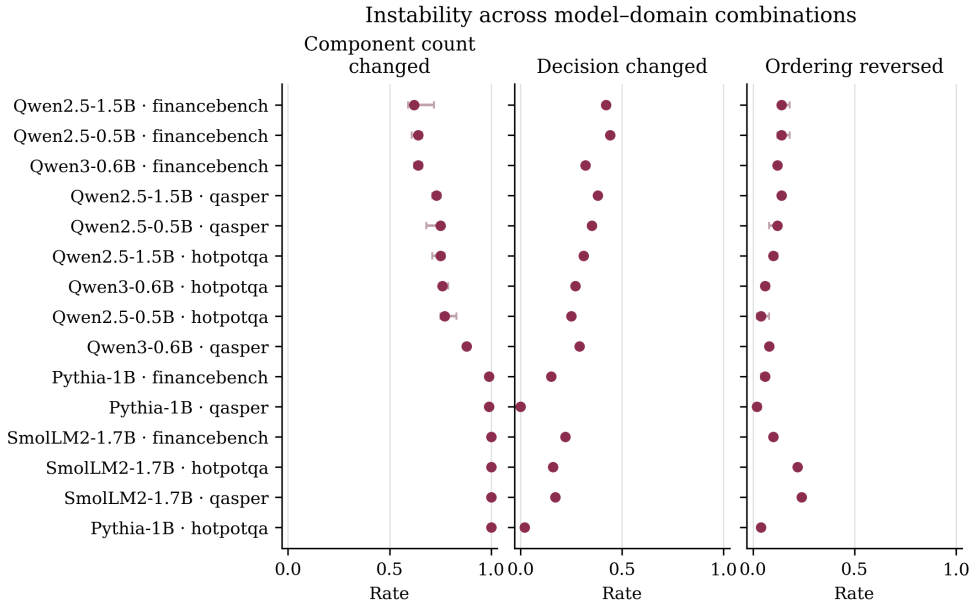

Fig. 1 Variability across the 15 model-domain combinations. The points indicate the median value across the five fixed transformations. The bars show the minima and maxima across the five transformations. The bars indicate variability across the five evaluated transformations rather than a confidence interval across a population of potential transformations.

0.880]. There was higher disagreement in reparameterisation for all 12 combinations. This shows that changing the basis has a much larger effect on component selection than changing the permutation seed. However this is just a descriptive interval as the combinations share models and domains. The median decision disagreement rate is 0.30 in reparameterisation. As the scores were not saved in the seed run there is not a decision level comparison. Figure 3 summarises the seed and preprocessing controls.

Table 5 Post-hoc controls across 12 model–domain combinations. Rates are fractions of 100 evaluation contexts. The reparameterization rate is summarized over five transformations for the seed comparison. The seed control uses one independent PA rerun and reports component-count disagreement.

Quantity	Median
Reparameterization component-count disagreement	0.76
No-reparameterization independent-seed rate	0.06
Paired reparameterization-minus-seed difference	0.69
Reparameterization decision disagreement	0.30
Centring-only reparameterization rate	1.00
Standardized-minus-centring-only difference	−0.24

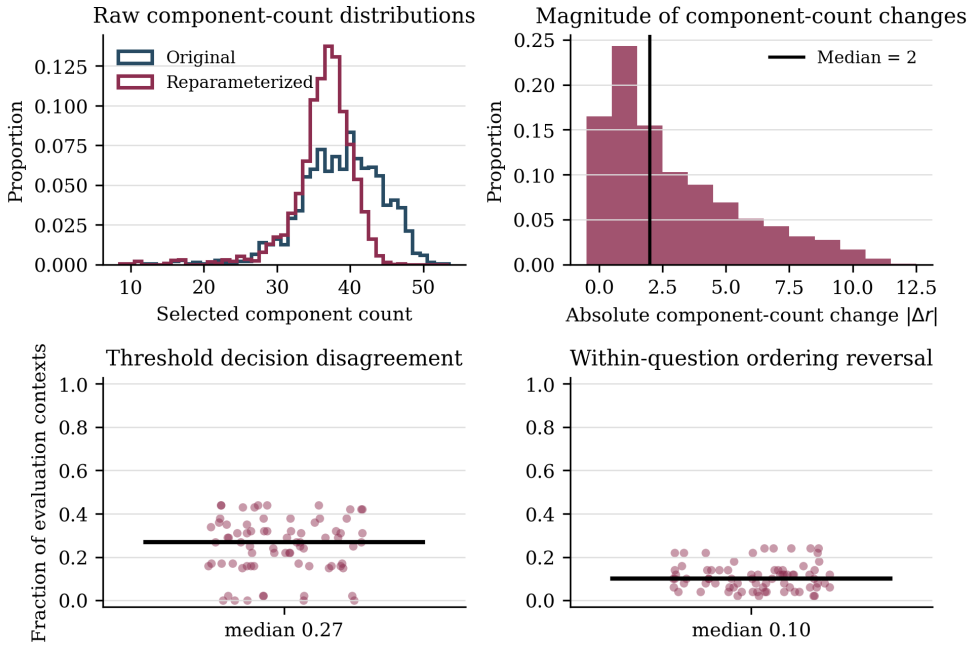

Fig. 2 Component-count changes, decision disagreement, and ordering reversals. **Top left,** raw selected component counts for 1,500 original contexts and 7,500 matched reparameterized contexts. **Top right,** absolute changes for the 7,500 matched context transformations. **Bottom,** decision disagreement and within-question ordering reversal for each of the 75 model–domain–transformation comparisons. Horizontal segments mark medians. Decisions use the score threshold selected on the original calibration split.

The second control removes standardization. The main scoring scheme performs standardization on each coordinate. However, this depends on the basis; both the observed spectrum and reference spectrum will be affected by standardization. In the centring-only version the data are centred but not standardized, so an orthogonal change of basis leaves the observed spectrum unchanged. This implies that any change in the count must come from the reference step. This is the setting of Lemma 1 and the PA count results in Theorems 1 and 2. If X_c denotes the centered context matrix, then the reparameterized centered matrix is exactly X_cR, and

$$(X_cR)^\top(X_cR) = R^\top X_c^\top X_cR.$$

It follows that the observed eigenvalues are identical in exact arithmetic. Any change in the selected component count must arise from the column-permutation reference distribution, apart from numerical error. For the observed covariance spectra, the saved control summary reports a maximum spectrum-normalized discrepancy of 1.26×10^{-6},

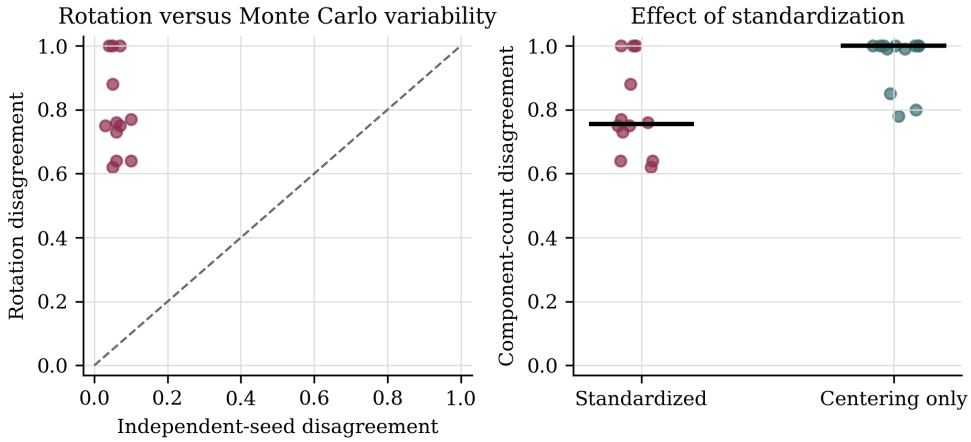

Fig. 3 Controls across 12 model–domain combinations. The left panel shows reparameterization-induced component-count disagreement versus disagreement from an independent PA seed without reparameterization, and the dashed line denotes equality. The right panel shows component-count disagreement under standardized and centring-only preprocessing. Horizontal segments mark medians.

using

$$\frac{\max_k |\lambda'_k - \lambda_k|}{\max\{\max_k |\lambda_k|, 10^{-12}\}}.$$

This is a spectrum-normalized absolute error, not the maximum individual relative eigenvalue error.

Although the observed spectrum is unchanged, the number of selected components varies in 1,141 out of 1,200 contexts, and the median of the model–domain disagreement rates is 1.00 (range 0.78–1.00). The median of the standardized minus centring-only difference is -0.24 (model–domain bootstrap-based 95% confidence interval $[-0.310, 0.025]$). Therefore, coordinate-wise standardization is not needed to induce instability.

The standardized analysis has two additional sources of change. Let $\widetilde{X} = X_c D_X^{-1}$ use the implementation's $n-1$-denominator sample standard deviations and let $\widetilde{S} = \widetilde{X}^\top \widetilde{X}/n$. For coordinates whose sample standard deviations are at least the numerical floor, the diagonal entries equal $c_n = (n-1)/n$. If the floor does not alter any coordinate's standard deviation, Lemma 1 gives

$$\mathbb{E}_\pi[\widetilde{S}^\pi] = c_n I_d, \qquad V(\widetilde{S}) = d(d-1)c_n^2.$$

These identities hold for all coordinate systems where the floor doesn't change the standard deviation of any coordinate. Standard deviation with $1/n$ denominator simplifies the expressions to I_d and $d^2 - d$ respectively. Standardisation therefore ensures that the expected reference covariance will have the same diagonal across all bases. Whilst this removes one source of basis dependence, the scoring procedure still depends upon the basis. First, standardising and then changing the basis does not give the same result as changing the basis and then standardising. Therefore, there could be

differences in the observed standardised spectrum. Second, the finite permutation reference may vary further than expected due to it having coordinate dependent higher moments. Therefore, using the controls we see that neither maintaining the observed spectrum nor balancing coordinate variances ensures stability of the PA counts. The centring-only control shows that the reference procedure isn't stable when maintaining the observed spectrum. The standardisation ensures that coordinate variances are balanced but the complete scoring procedure depends on the basis.

4.4 Context Discrimination

Stability in this instance is only useful if we have a score that can differentiate between contexts with and without annotated evidence. Table 6 shows the held-out AUROC for the 15 model domain combinations. We can see that the ridge score gives the highest median AUROC (0.712) compared to the PA score (0.703). Ridge outperforms PA in 11/15 matched pairs with a median difference of 0.028. A bootstrap descriptive 95% interval was calculated for the model domain pairs with the result being $[-0.003, 0.051]$. Zero falls within this interval so we cannot conclude that ridge has an advantage in prediction. However, the invariant scores do provide information about the conditions of the context whilst providing stability. Figure 4 shows the comparison in discrimination and sensitivity to reparameterisation.

Table 6 Held-out context discrimination across 15 model–domain combinations. Ranges show the minimum and maximum AUROC across combinations.

Score	Median AUROC	Range
PA projection score	0.703	[0.637, 0.813]
Variance-threshold score	0.697	[0.664, 0.824]
Ridge score	0.712	[0.683, 0.827]
Answer log-probability	0.689	[0.527, 0.839]

4.5 Robustness Analysis

The independent seed control has a considerably lower component-count disagreement rate. A sensitivity analysis was performed to investigate whether the high component-count disagreement rate under a change of basis is present under different PA settings. We consider one representative model and domain combination for each of the four architecture families with one fixed function-preserving transformation. Each combination had 100 held-out contexts.

We varied the number of permutation replicates (50, 100, 200), the reference quantile (0.90, 0.95, 0.99) and the selection rule. The selection rule can either select consecutive components from the largest until the first failure or count all components above their reference thresholds. The counts were considered separately at the three measured layers and the median. This gives 4 options for reporting the number of

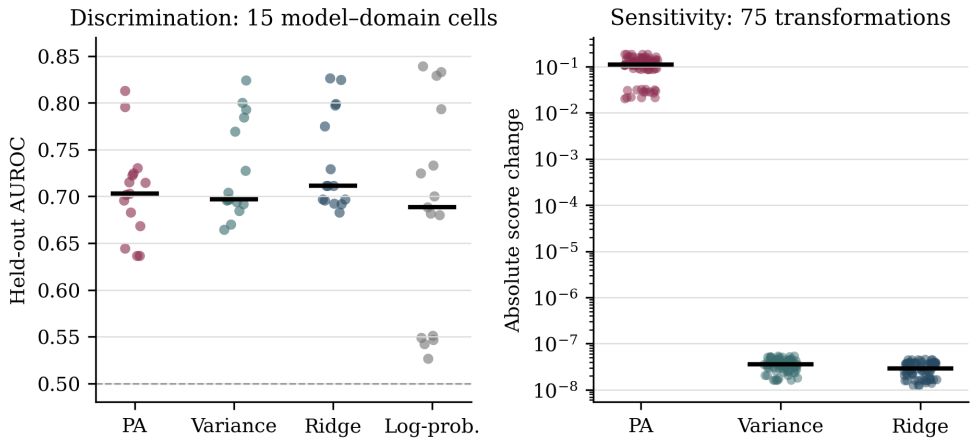

Fig. 4 Left panel. Held-out AUROC for the PA projection, variance-threshold, ridge, and mean answer log-probability scores. Each point is one of 15 model–domain combinations, and each horizontal segment is the median. **Right panel.** Sensitivity to function-preserving reparameterization. Each point is one of 75 model–domain–transformation comparisons. For PA, points show the median absolute score change across evaluation contexts. For the invariant scores, they show the maximum absolute change. The latter gives a conservative check of numerical stability.

components. This gives 72 total settings to consider in a $3 \times 3 \times 2 \times 4$ grid. For the smaller permutation budgets we used the first 50/100 replicates from the same 200 replicates. Across the grid the component-count disagreement rate remains high. In the original setting with 50 replicates, 0.95 quantile, consecutive selection and median layer counts, the median disagreement rate across the 4 representative combinations is 0.87 (question bootstrap 95% CI [0.825, 0.915]). Across all 72 settings the median disagreement rate ranges from 0.86 to 0.99. This indicates that there is sensitivity to the hidden basis in the number of components selected for these combinations and transformations. Table 7 and Figure 5 show comparisons varying one option at a time whilst keeping the others constant at the main setting.

Table 7 One-setting-at-a-time comparisons from the full sensitivity analysis in four representative model–domain settings. Each row varies one choice while holding the others at 50 replicates, the 0.95 permutation-reference quantile, consecutive-from-largest selection, and median-across-layer aggregation.

Choice	Values	Median rate
Permutation replicates	50 / 100 / 200	0.870 / 0.865 / 0.865
Reference quantile	0.90 / 0.95 / 0.99	0.880 / 0.870 / 0.885
Selection rule	Consecutive from largest / count all	0.870 / 0.870
Layer aggregation	Quarter / half / three-quarter / median	0.925 / 0.980 / 0.935 / 0.870

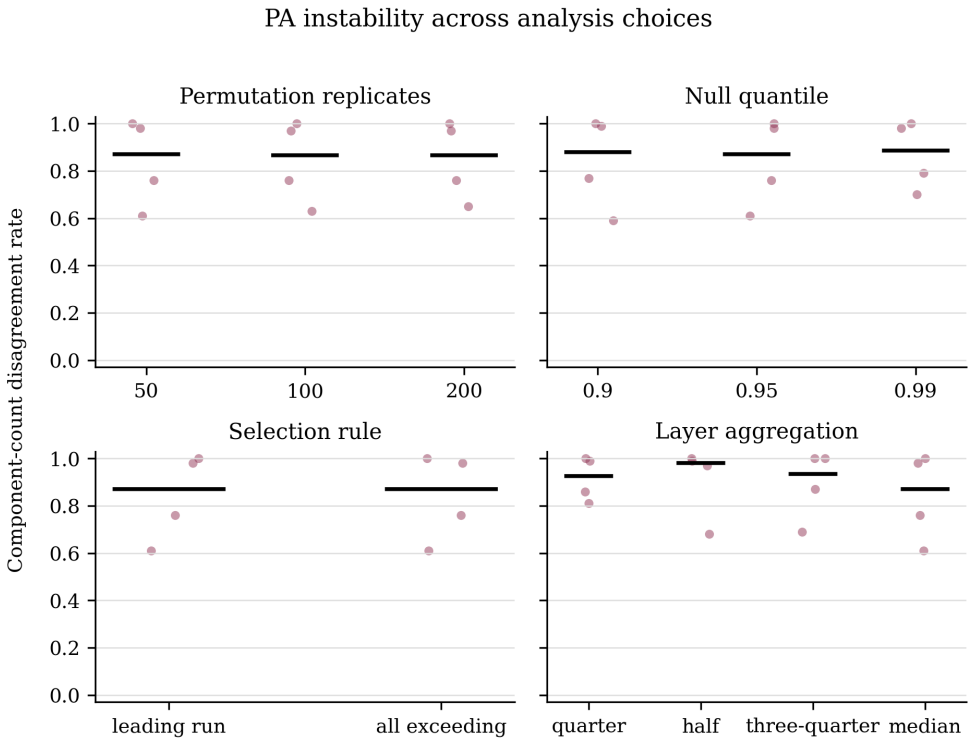

Fig. 5 Context-level PA component-count disagreement rates across analysis choices. Points are the four representative model–domain–transformation settings, and horizontal segments mark medians. In each panel, settings not shown are held at the main experimental setting.

5 Discussion and Limitations

We find that PA can be sensitive to the choice of hidden basis, even though the change of basis leaves the mathematical input-output function of the model unchanged. This is caused by the reference in PA preserving the coordinate empirical distributions, which depend on the basis. Theoretical results show the number of components can be sensitive, and empirical results show sensitivity can also occur in projections and classifications.

The results show that the number of components can be sensitive to the reparameterisation. There is a large increase in the number of component disagreements compared to the independent seed control. The number of disagreements is large for all the tested PA settings. However, the independent seed control uses only one run per combination. The Gram based scores are invariant numerically, and were able to achieve similar results to PA in terms of held out discrimination. This shows that information about the context conditions can still be obtained without sensitivity to the basis. However, the AUROC results do not indicate any improvements in prediction. In the centring-only setting, the selected subspace transforms with the basis, so when the number of components chosen per layer is fixed and the cutoff does not split

a repeated eigenvalue, the projection score is unchanged. Therefore, in that setting, for the score or classification to be sensitive there must be a change in the number of components. In the standardised results, both the observed spectrum and reference can change. Therefore it is not shown that the number of components being chosen is the cause of the sensitivity. The general results for the reference use the full orthogonal group. The LayerNorm results use a smaller subgroup that keeps the mean direction invariant.

The experiments look at column permutation PA for one task with retrieved context. This shows that the measure can be sensitive, but does not imply all representation based methods are unstable. The context creation can cause issues. The contexts with and without annotated evidence can have different lengths, and truncating to 192 tokens may cause loss of evidence. The analyses do not look at how the length affects the results or if evidence is retained. The AUROC results therefore only indicate that the scores are sensitive to the context conditions as constructed. However, the reparameterisation comparison is still valid as the original and transformed measures used the same input. Overall, we can see that there is a clear need to consider the interpretation of the measure. If a measure is claimed to be a property of the model, it should be checked under transformations that are compatible with the architecture, and keep the function invariant. If the measure is sensitive then the interpretation should take into account the choice of hidden basis.

6 Conclusion

This study shows that PA can be sensitive to properties of the chosen hidden basis. An orthogonal change of basis does not change the covariance spectrum, but the number of components chosen by column permutation PA can change. This is due to the reference preserving the coordinate empirical distributions, which depend on the basis. The general theorem shows that there is a conflict between preserving the marginal distributions exactly, the reference being fully orthogonally equivariant, and a change in the empirical second moment matrix. The results show that this conflict can occur in PA causing sensitivity in the number of components. The experiments show that across the tested model, domain and transformation combinations there is a median disagreement rate of 0.79 for the number of components and 0.27 for the classifications. The number of components changes far more often than in the independent seed control. The centring-only results show that there is sensitivity in the number of components while keeping the observed covariance spectrum invariant. In the standardised results, both the spectrum and reference can change so it cannot be said that the number of components being chosen is the cause of the sensitivity. The invariant comparison scores show that it is possible to have a measure that is not sensitive to the basis. The Gram based scores changed only at the level of numerical error and achieved similar AUROC results to PA. While this does not show any improvements in prediction, it does show that it is possible to achieve useful discrimination. While the measure may be informative, the interpretation should specify what changes in the representation should not affect the measure. Preserving the coordinate distributions does not make the number of components selected by PA invariant. Generally,

if a claim is made about the dimensionality/behaviour of the model, evidence should be shown that the measure is invariant to transformations that are compatible with the architecture and preserve the model function. If it is not then the hidden basis is implicitly part of what is being measured.

Declarations

- **Funding:** N/A
- **Competing interests:** No competing interest.
- **Ethics approval and consent to participate:** N/A
- **Consent for publication:** N/A
- **Code and data availability:** The code, experimental protocols, and materials required to reproduce the analyses, tables, and figures are available at https://github.com/abdullahkarasan/representation-measurement-invariance.
- **Author contribution:** The sole author carried out all parts of the work.

Appendix A Proofs of the Main Technical Results

This appendix collects the longer proofs omitted from the main presentation.

A.1 Proof of Lemma 2

Proof For $j \neq k$, the (j,k) entry of S^{π} has the same distribution as $x_j^{\top} P x_k / n$, where P is uniform. Sampling without replacement and column centering give

$$\mathbb{E}_{\pi}\left[\left(\frac{x_j^{\top} P x_k}{n}\right)^2\right] = \frac{S_{jj} S_{kk}}{n-1}.$$

The diagonal of S^{π} equals that of S deterministically. Summing the off-diagonal second moments proves the identity. Since $\|M\|_{\mathrm{op}} \leq \|M\|_{\mathrm{F}}$, Markov's inequality applied to the squared Frobenius norm proves the probability bound. Weyl's inequality gives the simultaneous eigenvalue statement. Taking $\delta = \alpha$ under the quantile convention in Lemma 2 gives the threshold bound. □

A.2 Proof of Theorem 1

Proof Choose an orthogonal matrix R_{loc} whose first coordinate is a leading eigenvector of S. The first diagonal entry of $R_{\mathrm{loc}}^{\top} S R_{\mathrm{loc}}$ is then $\lambda_1(S)$. Column permutation preserves every covariance diagonal entry exactly, so every reference realization has largest eigenvalue at least $\lambda_1(S)$. Consequently $t_1(X R_{\mathrm{loc}}) \geq \lambda_1(S)$, the strict first comparison fails, and $\widehat{r}_{\mathrm{PA}}(X R_{\mathrm{loc}}) = 0$.

By the Schur–Horn theorem, there is also an orthogonal matrix R_{flat} for which every diagonal entry of $R_{\mathrm{flat}}^{\top} S R_{\mathrm{flat}}$ equals $\bar{\lambda} = \mathrm{tr}(S)/d$. In this basis,

$$V(R_{\mathrm{flat}}^{\top} S R_{\mathrm{flat}}) = (\mathrm{tr}\, S)^2 - \frac{(\mathrm{tr}\, S)^2}{d}.$$

Lemma 2 therefore gives $t_k(X R_{\mathrm{flat}}) \leq c_{\alpha}(S,n)$ simultaneously for every componentwise threshold. The observed eigenvalues remain $\lambda_k(S)$. By the definition of $m_{\alpha}(S,n)$, the first $m_{\alpha}(S,n)$ observed eigenvalues strictly exceed their thresholds, so $\widehat{r}_{\mathrm{PA}}(X R_{\mathrm{flat}}) \geq m_{\alpha}(S,n)$. Comparing the two bases proves the result. □

A.3 Proof of Theorem 2

Proof Both observed covariance matrices have spectrum $(\lambda, 0, \ldots, 0)$. Moreover, $\|e_1\| = \|w\| = 1$ and $\mathbf{1}^\top e_1 = \mathbf{1}^\top w = 1$. An orthogonal transformation that fixes $\mathbf{1}$ can map any such unit vector to another with the same inner product with $\mathbf{1}$. There is therefore an $R \in \mathcal{G}_{\mathrm{LN}}$ with $R^\top e_1 = w$.

For X_{loc}, every reference realization has the same spectrum because its only nonzero column retains squared norm $n\lambda$. Thus $t_1 = \lambda$, and the strict comparison fails.

For X_{mix}, the largest covariance diagonal entry is $4\lambda/9$, and

$$V(S) = \lambda^2 \left(1 - \sum_{j=1}^{d} w_j^4\right) = \frac{16}{27}\lambda^2.$$

Lemma 2 gives

$$t_1 \le \frac{4\lambda}{9} + \lambda\sqrt{\frac{16}{27(n-1)\alpha}}.$$

The sample-size condition makes this bound strictly smaller than λ, so the first component is retained. The second observed eigenvalue is zero and its reference threshold is nonnegative, so the procedure stops after one component. □

A.4 Proof of Theorem 3

Proof Fix a unit vector v and choose $R \in O(d)$ such that $Re_1 = v$. Exact marginal preservation applied to XR implies that the first projected empirical measure of a draw from $\mathcal{K}(XR, \cdot)$ equals $\widehat{\mu}_{Xv}$ almost surely. By right-orthogonal equivariance,

$$\widehat{\mu}_{Zv} = \widehat{\mu}_{Xv}$$

for $Z \sim \mathcal{K}(X, \cdot)$. Apply the argument on a countable dense subset of the unit sphere and intersect the corresponding probability-one events. Continuity in v extends the identity to every direction. The Cramér–Wold theorem then yields $\widehat{\mu}_Z = \widehat{\mu}_X$. Equality of finite empirical measures with equal atom weights implies equality of their row multisets, hence $Z = PX$. □

Appendix B Function-Preserving Parameter Maps

The construction acts on residual states written as row vectors. For an affine input projection $hW^\top + b$, replacing h by hR and W by WR leaves the projection unchanged. For an affine output projection into the residual stream, use $W' = R^\top W$ and $b' = bR$, so its output is right-multiplied by R. Token embeddings transform as $E' = ER$. The same transformation applies to residual additions because $(h + f(h))R = hR + f(h)R$.

For RMSNorm, write the normalization as $\nu(h)D_g$, where $\nu(h) = h/\sqrt{d^{-1}\|h\|^2 + \epsilon_{\mathrm{norm}}}$. Multiply the coordinatewise normalization gains into each subsequent linear map using $W \leftarrow WD_g$, then set the normalization gain to one. Since $\nu(hR) = \nu(h)R$ for $R \in O(d)$, the remaining normalization commutes with the residual rotation. Query, key, value, gate, and up projections use the input map; attention-output and down projections use the output map. The resulting query and key values are unchanged, so rotary position embeddings and Qwen3's query/key normalizations need no residual coordinate transformation of their own. The final normalization gain is absorbed into the output head before applying its input map.

If input embeddings and the output head are tied, absorbing the final gain into the output head generally requires an independent output-head parameter. The implementation clones that parameter before fusion and disables weight tying. The transformed decoder computes the same function, but it need not retain the original weight-sharing constraint.

For LayerNorm, let $C = I - \mathbf{1}\mathbf{1}^\top/d$. The subgroup $R\mathbf{1} = \mathbf{1}$ satisfies $CR = RC$ and preserves centered norms. If an affine projection consumes $\nu_C(h)D_g + \beta$, absorb the gain and offset as

$$W \leftarrow WD_g, \qquad b \leftarrow b + \beta W_{\text{old}}^\top,$$

and set the LayerNorm gain and offset to one and zero, respectively. Here $\nu_C(h) = hC/\sqrt{d^{-1}\|hC\|^2 + \epsilon_{\text{norm}}}$. The same input and output maps then apply to GPT-NeoX attention and MLP projections, embeddings, and output head. Any bias parameters created by this adjustment are included in the transformed decoder. The measured residual states precede the final normalization; the returned post-normalization states need not satisfy the same $h' = hR$ relation after gain fusion.

All these equalities are exact-arithmetic statements. The actual stored activations and matrix products applied after extraction are stored or computed in float32; the Gram-based comparison scores convert their inputs to float64. Numerical checkpoint agreement is a separate empirical question.

Appendix C Numerical Checkpoint Verification

The main analysis applies $X \mapsto XR$ directly to stored representations. The checkpoint checks separately test whether the parameter maps produce the expected hidden states while preserving model outputs numerically.

Each model–seed pair is tested on ten fixed, teacher-forced prompts. A pair passes if the relative logit error $\|L' - L\|_2 / \max\{\|L\|_2, 10^{-12}\}$ and the maximum prompt-level change in mean sequence log-probability are both at most 10^{-4}, with identical next-token argmax predictions. Hidden-state errors are measured against $H_\ell R$ at the three studied layers.

Table C1 summarizes 22 of the 25 pairs, of which 21 pass. Qwen2.5-1.5B at seed 303 slightly exceeds the logit tolerance, with an error of 1.0437×10^{-4}, despite unchanged argmax predictions. SmolLM2 seeds 303, 404, and 505 remain unchecked. The records document checkpoint revisions, prompt tokens, software, precision, and execution settings.

Two additional archived Qwen2.5-1.5B seed-101 runs also exceed the tolerance, with relative logit errors of 1.1081×10^{-4} and 4.2749×10^{-4} and unchanged argmax predictions. Thus, the checks support the parameter maps across most tested pairs but do not establish uniform numerical agreement. They cover neither higher precision nor free-running generation.

The stored-state analysis includes all fixed transformations without applying the protocol's checkpoint-level exclusion rule. Its findings therefore concern measurement dependence under the mathematical changes of basis. They do not require, or establish, identical finite-precision checkpoint outputs on every evaluation context.

Table C1 Float32 checkpoint verification. Checked pairs are reported out of five seeds, and passes are reported out of completed checks. Errors are maxima across completed checks. Logit and hidden-state errors are relative ℓ_2 errors. Mean logp is the maximum absolute change in prompt-level mean sequence log-probability.

Model	Checked	Passed	Logits	Mean logp	Hidden
Qwen2.5-0.5B	5/5	5/5	3.80×10^{-6}	4.77×10^{-6}	9.28×10^{-7}
Qwen2.5-1.5B	5/5	4/5	1.04×10^{-4}	4.89×10^{-6}	2.46×10^{-6}
Qwen3-0.6B	5/5	5/5	1.01×10^{-5}	3.34×10^{-6}	1.20×10^{-6}
SmolLM2-1.7B	2/5	2/2	4.85×10^{-6}	4.53×10^{-6}	7.20×10^{-7}
Pythia-1B	5/5	5/5	1.18×10^{-6}	2.38×10^{-6}	3.13×10^{-7}

References

Antverg, O., Belinkov, Y.: On the pitfalls of analyzing individual neurons in language models. In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=8uz0EWPQIMu

Ainsworth, S.K., Hayase, J., Srinivasa, S.S.: Git re-basin: Merging models modulo permutation symmetries. In: International Conference on Learning Representations (2023). https://openreview.net/forum?id=CQsmMYmlP5T

Ashkboos, S., Mohtashami, A., Croci, M.L., Li, B., Jaggi, M., Alistarh, D., Hoefler, T., Hensman, J.: QuaRot: Outlier-free 4-bit inference in rotated LLMs. In: Advances in Neural Information Processing Systems, vol. 37 (2024). https://papers.nips.cc/paper_files/paper/2024/hash/b5b939436789f76f08b9d0da5e81af7c-Abstract-Conference.html

Ben Allal, L., Lozhkov, A., Bakouch, E., Martín Blázquez, G., Penedo, G., Tunstall, L., Marafioti, A., Kydlíček, H., Piqueres Lajarín, A., Srivastav, V., et al.: SmolLM2: When Smol Goes Big—Data-Centric Training of a Small Language Model (2025)

Buja, A., Eyuboglu, N.: Remarks on parallel analysis. Multivariate Behavioral Research **27**(4), 509–540 (1992) https://doi.org/10.1207/s15327906mbr2704_2

Biderman, S., Schoelkopf, H., Anthony, Q.G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M.A., Purohit, S., Prashanth, U.S., Raff, E., Skowron, A., Sutawika, L., Wal, O.: Pythia: A suite for analyzing large language models across training and scaling. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 2397–2430. PMLR, ??? (2023)

Cheng, E., Doimo, D., Kervadec, C., Macocco, I., Yu, L., Laio, A., Baroni, M.: Emergence of a high-dimensional abstraction phase in language transformers. In: International Conference on Learning Representations (2025). https://openreview.net/forum?id=0fD3iIBhlV

Ding, F., Denain, J.-S., Steinhardt, J.: Grounding representation similarity through statistical testing. In: Advances in Neural Information Processing Systems, vol. 34, pp. 1556–1568 (2021). https://proceedings.neurips.cc/paper/2021/hash/0c0bf917c7942b5a08df71f9da626f97-Abstract.html

Dasigi, P., Lo, K., Beltagy, I., Cohan, A., Smith, N.A., Gardner, M.: A dataset of information-seeking questions and answers anchored in research papers. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4599–4610. Association for Computational Linguistics, ??? (2021). https://doi.org/10.18653/v1/2021.naacl-main.365

Dobriban, E., Owen, A.B.: Deterministic parallel analysis: An improved method for selecting factors and principal components. Journal of the Royal Statistical Society: Series B **81**(1), 163–183 (2019) https://doi.org/10.1111/rssb.12301

Dobriban, E.: Permutation methods for factor analysis and PCA. The Annals of Statistics **48**(5), 2824–2847 (2020) https://doi.org/10.1214/19-AOS1907

Dinh, L., Pascanu, R., Bengio, S., Bengio, Y.: Sharp minima can generalize for deep nets. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1019–1028. PMLR, ??? (2017)

Godfrey, C., Brown, D., Emerson, T., Kvinge, H.: On the symmetries of deep learning models and their internal representations. In: Advances in Neural Information Processing Systems, vol. 35 (2022). https://proceedings.neurips.cc/paper_files/paper/2022/hash/4df3510ad02a86d69dc32388d91606f8-Abstract-Conference.html

Hayne, L., Jung, H., Carter, R.M.: Does representation similarity capture function similarity? Transactions on Machine Learning Research (2024)

Hewitt, J., Liang, P.: 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, pp. 2733–2743. Association for Computational Linguistics, ??? (2019). https://doi.org/10.18653/v1/D19-1275

Horn, J.L.: A rationale and test for the number of factors in factor analysis. Psychometrika **30**(2), 179–185 (1965) https://doi.org/10.1007/BF02289447

Hong, D., Sheng, Y., Dobriban, E.: Selecting the number of components in PCA via random signflips. arXiv preprint arXiv:2012.02985 (2020) https://doi.org/10.48550/arXiv.2012.02985

Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., Vidgen, B.: FinanceBench: A New Benchmark for Financial Question Answering (2023)

Joren, H., Zhang, J., Ferng, C.-S., Juan, D.-C., Taly, A., Rashtchian, C.: Sufficient context: A new lens on retrieval-augmented generation systems. In: International Conference on Learning Representations (2025). https://openreview.net/forum?id=Jjr2Odj8DJ

Kriegeskorte, N., Mur, M., Bandettini, P.A.: Representational similarity analysis—connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience **2**, 4 (2008) https://doi.org/10.3389/neuro.06.004.2008

Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 3519–3529. PMLR, ??? (2019)

Klabunde, M., Wald, T., Schumacher, T., Maier-Hein, K., Strohmaier, M., Lemmerich, F.: ReSi: A comprehensive benchmark for representational similarity measures. In: International Conference on Learning Representations (2025). https://openreview.net/forum?id=PRvdO3nfFi

Leask, P., Bussmann, B., Pearce, M.T., Bloom, J.I., Tigges, C., Al Moubayed, N., Sharkey, L., Nanda, N.: Sparse autoencoders do not find canonical units of analysis. In: International Conference on Learning Representations (2025). https://openreview.net/forum?id=9ca9eHNrdH

Lee, J.H., Jiralerspong, T., Yu, L., Bengio, Y., Cheng, E.: Geometric signatures of compositionality across a language model's lifetime. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, pp. 5292–5320. Association for Computational Linguistics, ??? (2025). https://doi.org/10.18653/v1/2025.acl-long.265

Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., Chandra, V., Tian, Y., Blankevoort, T.: SpinQuant: LLM quantization with learned rotations. In: International Conference on Learning Representations (2025). https://openreview.net/forum?id=ogO6DGE6FZ

Nielsen, B.M.G., Marconato, E., Dittadi, A., Gresele, L.: When does closeness in distribution imply representational similarity? an identifiability perspective. In: Advances in Neural Information Processing Systems, vol. 38 (2025). https://doi.org/10.52202/085713-3418

Pimentel, T., Valvoda, J., Maudslay, R.H., Zmigrod, R., Williams, A., Cotterell, R.: Information-theoretic probing for linguistic structure. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4609–4622. Association for Computational Linguistics, ??? (2020). https://doi.org/10.18653/

v1/2020.acl-main.420

Qwen Team: Qwen2.5 Technical Report (2024)

Qwen Team: Qwen3 Technical Report (2025)

Roeder, G., Metz, L., Kingma, D.P.: On linear identifiability of learned representations. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 9030–9039. PMLR, ??? (2021)

Ru, D., Qiu, L., Hu, X., Zhang, T., Shi, P., Chang, S., Jiayang, C., Wang, C., Sun, S., Li, H., Zhang, Z., Wang, B., Jiang, J., He, T., Wang, Z., Liu, P., Zhang, Y., Zhang, Z.: RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In: Advances in Neural Information Processing Systems, vol. 37 (2024). https://doi.org/10.52202/079017-0692

Saad-Falcon, J., Khattab, O., Potts, C., Zaharia, M.: ARES: An automated evaluation framework for retrieval-augmented generation systems. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 338–354. Association for Computational Linguistics, ??? (2024). https://doi.org/10.18653/v1/2024.naacl-long.20

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2369–2380. Association for Computational Linguistics, ??? (2018). https://doi.org/10.18653/v1/D18-1259