Gauge-Invariant Diagnostics for Hidden Representational Structure in Transformer Language Models

Anders Olsson[*]

February 2026

Abstract

We present a geometric framework for testing whether transformer language models contain internal state variation not determined by the output distribution. We formalize this using a fiber bundle perspective: the projection from hidden states to output distributions may be many-to-one, with fibers encoding path-dependent information. We derive gauge-invariant observables, classify computable quantities by their transformation properties, and define a four-criterion diagnostic for representational holonomy—path-dependent internal divergence that persists under matched output distributions. We apply the full diagnostic suite to Phi-2 (2.7B parameters) under a controlled path-manipulation protocol. All experiments return negative results: output-layer geometry reduces to entropy (Experiment 0), path dependence is visible in the output distribution, failing the matched-output condition required for holonomy (Experiment 1), and intermediate-layer divergence reflects subspace rotation rather than information compression (Experiment 2). The primary contribution is a falsifiable, reusable diagnostic framework for detecting hidden representational structure in autoregressive transformers.

1 Introduction

When a transformer language model processes a prompt, its internal state—a high-dimensional activation vector at each layer—determines a probability distribution over next tokens. A natural question arises: does the internal state carry information *beyond* what the output distribution reveals?

This question has practical implications for interpretability (are there "hidden" features that probing the output misses?), for alignment (can a model's internal state diverge from its expressed behavior?), and for the broader study of representational dynamics in deep networks. It also connects to longstanding questions in philosophy of mind about whether internal states have structure beyond their functional role, though we do not pursue that connection here.

We do not claim that representational holonomy is expected in current models; rather, we formalize the conditions under which it would exist and provide a reusable diagnostic for testing those conditions empirically.

We formalize the question geometrically. Let $\mathcal{Z}$ denote the manifold of hidden states and $\mathcal{Q}$ the simplex of output distributions. The model defines a smooth map $\varphi : \mathcal{Z} \to \mathcal{Q}$. If φ is injective (up to the relevant equivalence), the output determines the internal state—there is no hidden structure. If φ is many-to-one, the preimage $\varphi^{-1}(q)$ for a given output distribution q is a nontrivial manifold: the *fiber* over q. Different points in the same fiber produce identical predictions but carry different internal structure.

[*] Independent researcher. Correspondence: anders@kliv.dev

The fiber bundle perspective yields a specific, testable prediction: *representational holonomy.* If a model is driven through a sequence of contexts that returns to the same output distribution, the internal state may not return to its starting point. The residual displacement in the fiber is holonomy, and its detection requires:

1. Verified output equivalence (the projection to $\mathcal{Q}$ has converged).
2. Measurable internal divergence (the fiber coordinate has shifted).

In this paper, we develop the mathematical framework (§2), classify observables by their gauge-invariance properties (§3), define a four-criterion diagnostic for holonomy (§4), report three experiments on Phi-2 (§5), and discuss what a positive result would require (§6).

2 Geometric Framework

2.1 Spaces and Maps

Let $\mathcal{Z} \subseteq \mathbb{R}^d$ be the activation space at a given layer, and let $\mathcal{Q} = \Delta^{V-1}$ be the probability simplex over a vocabulary of size V. A transformer with fixed weights defines a differentiable map $\varphi : \mathcal{Z} \to \mathcal{Q}$ that sends an activation vector z to the output distribution $p = \varphi(z) = \mathrm{softmax}(Uz + b)$ at the final layer (where $U \in \mathbb{R}^{V\times d}$ is the unembedding matrix and $b \in \mathbb{R}^V$ the bias), or more generally through the composition of all remaining layers.

Definition 1 (Observable and rich states)**.** *The* observable state *is* $Q_{\text{obs}} = \varphi(z) \in \mathcal{Q}$*. The* rich state *is the full activation* $Q_{\text{rich}} = z \in \mathcal{Z}$*.*

When φ is not injective, Q_{rich} carries strictly more information than Q_{obs}. The fiber $\varphi^{-1}(q)$ parameterizes the hidden degrees of freedom.

2.2 Pullback Metric

The Fisher–Rao metric on $\mathcal{Q}$ is the unique Riemannian metric (up to scale) that is monotone under Markov morphisms [Čencov, 1982]. In coordinates on the simplex interior, $F_{ij}(p) = \sum_k \frac{1}{p_k}\frac{\partial p_k}{\partial \theta^i}\frac{\partial p_k}{\partial \theta^j}$. In logit coordinates $\ell = \log p$ (up to a constant), the Fisher metric takes the form:

$$F(z) = \mathrm{diag}(p) - pp^\top \tag{1}$$

The *pullback metric* on $\mathcal{Z}$ is:

$$G(z) = J_\varphi(z)^\top \, F(\varphi(z)) \, J_\varphi(z) \tag{2}$$

where J_φ is the Jacobian of φ. $G(z)$ is positive semi-definite (PSD), typically rank-deficient since F has rank $V - 1$ and J_φ rarely has full column rank. $G(z)$ measures how much infinitesimal perturbations of the hidden state change the output distribution, weighted by the Fisher geometry.

At the final pre-logit layer, where $\varphi(z) = \mathrm{softmax}(Uz + b)$, the pullback simplifies to:

$$G(z) = U^\top(\mathrm{diag}(p) - pp^\top)U = \mathrm{Cov}_p(U) \tag{3}$$

where $\mathrm{Cov}_p(U)_{ij} = \sum_k p_k \, u_{ki}u_{kj} - (\sum_k p_k u_{ki})(\sum_k p_k u_{kj})$ and u_k are rows of the unembedding matrix U.

For intermediate layers ℓ, the pullback through the remaining network gives:

$$G^{(\ell)}(z^{(\ell)}) = J_{\ell\to\text{logits}}^\top \, (\mathrm{diag}(p) - pp^\top) \, J_{\ell\to\text{logits}} \tag{4}$$

where $J_{\ell\to\text{logits}}$ is the Jacobian of the composite map from layer ℓ activations to output logits.

2.3 Effective Dimensionality

A scalar summary of the local geometry is the *effective dimensionality*:

$$d_{\text{eff}}(z) = \exp\left(-\sum_i \hat{\lambda}_i \log \hat{\lambda}_i\right), \qquad \hat{\lambda}_i = \frac{\lambda_i}{\sum_j \lambda_j} \tag{5}$$

where $\{\lambda_i\}$ are the eigenvalues of $G(z)$ (or of the whitened metric $\hat{G}$; see §3). This is the exponential of the Von Neumann entropy of the normalized spectrum, ranging from 1 (all weight on one direction) to d (isotropic).

2.4 Fisher–Rao Geodesic Distance

The geodesic distance on the statistical manifold $\mathcal{Q}$ under the Fisher–Rao metric is:

$$d_{\text{FR}}(p, q) = 2 \arccos\left(\sum_k \sqrt{p_k\, q_k}\right) \tag{6}$$

This is the spherical distance between the square-root representations $\sqrt{p}$ and $\sqrt{q}$ on the unit sphere, and is invariant under sufficient statistics.

3 Gauge Invariance

A change of basis in activation space $z' = Az$ (for invertible $A \in \text{GL}(d)$) is a *gauge transformation*—it changes the coordinate description of the hidden state without altering the model's input-output behavior. Meaningful geometric quantities must be invariant under such transformations, or their transformation properties must be explicitly tracked.

3.1 Three-Tier Classification

We classify computable observables by their transformation behavior:

Tier 1 (GL(d)-invariant): Quantities computed entirely from output distributions. These include $d_{\text{FR}}(p, q)$, the probe divergence $H = \mathbb{E}_i[d_{\text{FR}}(p(C_A \oplus \delta_i), p(C_B \oplus \delta_i))]$, Shannon entropy $H(p)$, and behavioral signatures. No dependence on activation coordinates.

Tier 2 (GL(d)-invariant with reference): Spectral features of the *whitened* metric $\hat{G} = M^{-1/2} G\, M^{-1/2}$, where $M = \mathbb{E}_{\text{probe}}[G(z)]$ is the probe-averaged metric. Under $z' = Az$, both G and M transform by congruence ($G' = A^{-\top} G A^{-1}$), so $\hat{G}$ undergoes a similarity transform and its eigenvalues are GL(d)-invariant. Equivalently, the generalized eigenvalues of (G, M) are coordinate-free.

Tier 3 (chart-dependent): Raw eigenvalues of G, activation norms $\|z_A - z_B\|$, cosine distances between activations. Useful for within-model diagnostics but not meaningful across coordinate systems.

Proposition 1 (Whitened metric invariance). *Let $G, M \succ 0$ be SPD matrices in $\mathbb{R}^{d \times d}$. Under the congruence action $G \mapsto A^{-\top} G A^{-1}$, $M \mapsto A^{-\top} M A^{-1}$ for $A \in \text{GL}(d)$, the spectrum of $\hat{G} = M^{-1/2} G M^{-1/2}$ is invariant.*

Proof. The generalized eigenvalue problem $Gv = \lambda M v$ is equivalent to $\hat{G} w = \lambda w$ where $w = M^{1/2} v$. Under the congruence action, $G' = A^{-\top} G A^{-1}$ and $M' = A^{-\top} M A^{-1}$, so $G'v' = \lambda M'v'$ becomes $A^{-\top} G(A^{-1}v') = \lambda A^{-\top} M(A^{-1}v')$, hence $G(A^{-1}v') = \lambda M(A^{-1}v')$. Setting $v = A^{-1}v'$, this is the original eigenvalue problem with the same λ. □

3.2 Regularization

When M is singular, we compute generalized eigenvalues of (G, M) restricted to range(M), discarding directions where $\lambda_M < \varepsilon$. Adding $\lambda_0 I$ for regularization breaks $\mathrm{GL}(d)$-invariance since I does not co-transform under congruence. Valid alternatives: (a) pseudoinverse on the support of M, or (b) regularize with a co-transforming reference $M_{\mathrm{reg}} = \mathbb{E}_{\mathrm{probe}}[G] + \lambda_0 \cdot G_{\mathrm{ref}}$ where G_{ref} is the metric at a fixed reference distribution.

4 Diagnostic Toolkit

We define a four-criterion test for representational holonomy. A model *passes* if all four criteria are satisfied for a given path-manipulation protocol.

4.1 Protocol: The Bank Experiment

Construct two context paths that traverse different semantic domains before arriving at an ambiguous target:

Path A: [Financial context] → [Nature context] → [Washout] → "The bank"
Path B: [Nature context] → [Financial context] → [Washout] → "The bank"

The washout buffer (a topically neutral passage) is designed to allow output distributions to converge while potentially preserving internal divergence. A library of follow-up probes $\{\delta_i\}$ is appended to both paths to measure behavioral divergence.

4.2 Four Criteria

Criterion 1 (Output equivalence — FR gate)**.** *The Fisher–Rao distance between output distributions at the target token satisfies* $d_{\mathrm{FR}}(p_A, p_B) < \tau$ *for a pre-specified threshold* τ *(e.g.,* $\tau = 0.05$*). This verifies that the two paths have converged in* $\mathcal{Q}$*.*

Criterion 2 (Non-monotonic divergence profile)**.** *Extracting activations* $z_A^{(\ell)}, z_B^{(\ell)}$ *at each layer* ℓ *at the target token position, the cosine distance profile* $\mathrm{d}_{\cos^{(\ell)}=1-\cos(z_A^{(\ell)},z_B^{(\ell)})}$ *has a peak at some intermediate layer* ℓ^* *with* $\mathrm{d}_{\cos^{(\ell^*)}}/\mathrm{d}_{\cos^{(L)}{}_{>1.2}}$ *(peak-to-output ratio exceeds 1.2). This indicates the model builds and then compresses path-dependent structure. The threshold* 1.2 *is a heuristic chosen to exclude trivial fluctuations; it is not theoretically grounded. Note: this is a Tier 3 (chart-dependent) diagnostic; it is necessary but not sufficient.*

Criterion 3 (Probe accuracy drop)**.** *A linear classifier (logistic regression with leave-one-group-out cross-validation) trained to predict path identity (A vs. B) from layer-*ℓ *activations shows accuracy that* decreases *from intermediate layers to the output layer. Specifically,* $\mathrm{acc}(\ell^*) - \mathrm{acc}(L) > 0.05$*. Perfect probe accuracy at the output layer indicates the information has not been compressed out—it has merely been rotated. Note: in high-dimensional settings with small sample sizes, linear separability may be trivial; this criterion is therefore interpreted relative to its behavior* across layers*, not as absolute evidence of stored information.*

Criterion 4 (LM-head null-space storage)**.** *Let* z *be the final-layer activation at the target token, and let* $\hat{w}$ *be a fixed unit vector in the final-layer activation space that separates paths A and B. We define* $\hat{w}$ *as the normalized mean-difference direction:* $w = \mathbb{E}[z \mid A] - \mathbb{E}[z \mid B]$*,* $\hat{w} = w/\|w\|$*.*

Define the scrubbed activation $z' = z - (z^\top \hat{w})\hat{w}$ *and recompute the output distribution via* $p' = \mathrm{softmax}(Uz' + b)$*. If the scrubbing distance satisfies* $d_{\mathrm{FR}}(p, p') < 0.01$*, the path information is consistent with storage in an approximate functional null space of the LM head—present in* z *but not used for next-token prediction. This is the strongest evidence for hidden fiber structure.*

4.3 Verdict Logic

Criteria satisfied	Interpretation
All four	**Positive:** representational holonomy detected
1 + 2 + 3, not 4	Information compressed but LM head still reads residual
1 + 2, not 3	Information rotated, not compressed (metric artifact)
Not 1	Inconclusive: output loop not closed; holonomy not testable

4.4 Controls

Each experiment includes mandatory controls:

C1 (noise floor): Same context forwarded twice; establishes numerical baseline.

C2 (perturbation sensitivity): Trivially edited washout (e.g., "twelve" → "12"); measures sensitivity to irrelevant variation.

C3 (disambiguation): Target replaced with unambiguous form ("The bank (financial institution)"); tests whether signal is purely word-sense ambiguity.

C4 (signed interaction): Sense-check probe ("refers to a") measuring whether finance-first paths produce higher finance-completion log-odds than nature-first paths; tests directionality.

C5 (recency equalization): Identical 200-token neutral suffix appended after washout, before target; controls for recency effects.

5 Experiments

All experiments use Phi-2 (2.7B parameters; Javaheripi and Bubeck 2023) in half-precision on a single RTX 3080 (10GB). Code is available at `https://github.com/anderswrk/representational-holonomy`.

5.1 Experiment 0: Output-Layer Geometry vs. Entropy

Question: Does the output-head pullback metric $G(z) = \mathrm{Cov}_p(U)$ capture structure beyond scalar entropy $H(p)$?

Design: Compute d_{eff} and $H(p)$ across 60 prompts spanning nine categories (creative, reasoning, factual, noise, code, philosophical, etc.) using a top-K ($K = 512$) approximation for G.

Results: d_{eff} correlates strongly with $H(p)$ (Pearson $r = 0.936$, Spearman $\rho = 0.948$). A matched-entropy residual analysis shows statistically significant category structure ($t = 2.84$, $p < 0.01$): code prompts fall below prediction (constrained vocabulary), philosophical prompts above (semantic diversity). Effect size is $\sim$2% of d_{eff} range.

Control (decisive): Row-permuting the unembedding matrix U—destroying semantic relationships between embedding vectors while preserving spectral statistics—yields a residual pattern correlated $r = 0.87$ with the real pattern. The category structure is driven by generic spectral properties of U interacting with the probability mass distribution, not by meaningful geometric arrangement.

Verdict: Negative. d_{eff} at the output head $\approx$ entropy + spectral artifact. The output-layer pullback metric does not extract geometric structure beyond what $H(p)$ provides.

5.2 Experiment 1: Path Dependence (Bank Experiment)

Question: Does the model exhibit path-dependent internal structure that is invisible to the output distribution?

Design: 16 path pairs (4 financial × 4 nature paragraphs), formatting washout (∼130 tokens), target "The bank." FR gate on output distributions. 30 probes across four categories (finance, nature, neutral, minimal) measured via KV-cache reuse. Sense-check probe: "refers to a" with finance/nature log-odds scoring. Full control suite C1–C4.

Results:

- FR gate failed: no pairs passed at $d_{\mathrm{FR}} < 0.02, 0.05$, or 0.1. Fallback to bottom-25% ($d_{\mathrm{FR}} \leq 0.33$), yielding 4 gated pairs. **Criterion 1 not satisfied.**
- Probe divergence: $H = 1.65$ (mean d_{FR}), well above C1 noise floor (0.0002) and 2.3× the C2 perturbation baseline (0.72).
- Signed interaction (C4): finance-first paths consistently produce higher finance log-odds ($\Delta = 1.83$, bootstrap 95% CI $[1.10, 2.61]$). Directional, topic-selective path dependence confirmed.
- Disambiguation (C3): did not collapse the signal ($d_{\mathrm{FR}} = 1.38$ vs. 1.65).

Verdict: Inconclusive for holonomy. Genuine directional path dependence exists and survives washout. However, because output distributions are not matched (Criterion 1 fails), the loop in $\mathcal{Q}$ is not closed: the internal divergence is expected given different outputs, and the holonomy hypothesis is not testable under these conditions. The result demonstrates ordinary context sensitivity, not hidden structure.

5.3 Experiment 2: Intermediate Layer Divergence

Question: Does path-dependent information peak at intermediate layers and get compressed toward the output?

Design: Forward all 16 Bank pairs with `output_hidden_states=True`. Extract activations at the target token position for all 33 layers (embedding + 32 transformer blocks). Compute cosine distance and L2 distance between path A and path B activations at each layer. While cosine distance is a Tier 3 (chart-dependent) diagnostic, it is valid for comparing layers within a single model checkpoint, where the activation basis is fixed. Linear probe (logistic regression, leave-one-pair-out CV, $C \in \{0.01, 0.1\}$) at each layer. LM-head projection test at the final layer. Five controls (A–D plus recency equalization).

Results:

Primary profile: Cosine distance peaks at layer 16 ($d_{\cos=0.0285}$), drops to 0.0118 at layer 32. Peak/output ratio 2.40×. Smooth, consistent across all 16 pairs. **Criterion 2 initially satisfied.**

Controls A–C (architectural controls):

Condition	Peak/output ratio	Interpretation
Bank (path-order)	2.40×	Large mid-network bulge
Different topics, no path swap	1.13×	Nearly flat
Same topic, different paragraphs	1.00×	Completely flat
Sentence-shuffled, path swap preserved	2.12×	Order sensitivity, not semantics

The bulge is specific to path-order manipulation (larger than controls A and B) but is driven by sequential processing order rather than semantic content (control C nearly matches).

Control D (linear probe): Classification accuracy = 1.000 at every layer from 1 to 32. Zero drop from intermediate to output layers. **Criterion 3 not satisfied.** However, this result must be interpreted cautiously: with $n = 32$ samples in $d = 2560$ dimensions, linear separability is expected for arbitrary label assignments (Cover's function counting theorem; Cover 1965). The diagnostic value lies not in the absolute accuracy but in its *constancy across layers*—if information were being compressed, accuracy would degrade toward the output even in the overparameterized regime. The flat profile, combined with the cosine bulge, confirms subspace rotation without information loss.

LM-head projection: Removing the probe's separating direction from layer-32 activations changes output distributions by $d_{\mathrm{FR}} = 0.113$. **Criterion 4 not satisfied.** The LM head actively reads path information.

Recency control (decisive): Adding a 200-token identical suffix before the target reduces the peak/output ratio from 2.40× to 1.22×—barely above the 1.2 heuristic threshold. The intermediate bulge is largely explained by distance from the divergent tokens in the sequence, not by depth of semantic processing. This implies that the "processing depth" reflected in the layer profile is primarily a function of token recency, not abstraction level.

Verdict: Negative. The non-monotonic cosine profile is real and path-specific, but reflects geometric reorganization (subspace rotation), not information compression. Path identity is perfectly preserved and functionally utilized at every layer. Phi-2 does not compress away path information between intermediate layers and output.

5.4 Summary of Results

	Criterion 1 FR gate	**Criterion 2** Profile bulge	**Criterio** Probe d
Required for holonomy	$d_{\mathrm{FR}} < \tau$	ratio > 1.2	acc drop >
Phi-2 result	FAIL (0.33)	PASS raw (2.40×); WEAK PASS after recency ctrl (1.22×)	FAIL (0.00

Phi-2 fails three of four criteria. No hidden representational holonomy is detected.

6 Discussion

6.1 What Was Found

Despite negative results for the holonomy hypothesis, the experiments revealed structured representational dynamics in Phi-2:

- **Persistent path dependence:** Context ordering produces consistent, directional effects on output distributions that survive >130 tokens of washout. Finance-first paths produce higher finance-completion probabilities even after extensive neutral buffering.
- **Layer-specific geometric reorganization:** Path-dependent representations undergo maximal angular divergence at mid-network layers before being rotated into different subspaces at the output. This reorganization does not destroy information.
- **Full information retention:** A linear probe recovers path identity perfectly at every layer. The model never forgets which context came first—it reorganizes the information rather than discarding it.

These findings are consistent with the residual stream acting as an information-preserving conduit [Elhage et al., 2021], with later layers selecting task-relevant features from a richer intermediate representation.

6.2 Why Phi-2 May Be the Wrong Test Subject

One possible explanation for the negative results is that in this model, path-dependent information remains functionally relevant at the output layer rather than migrating into an unused subspace. In a model where all representational dimensions contribute to the prediction task, there may be no room for functionally inert directions that could store hidden path information.

Concretely, a model would need to:

1. Build path-dependent representations at intermediate layers (observed in Phi-2).
2. Actively compress away that information in later layers, producing matched output distributions (not observed).
3. Retain the compressed information in dimensions the LM head does not utilize (not observed).

6.3 Connections to Related Work

The pullback metric construction connects to work on representation geometry in neural networks [Zavatone-Veth et al., 2023], information geometry of statistical models [Amari, 2016], and the study of how information flows through transformer layers [Voita et al., 2019, Geva et al., 2023]. The fiber bundle formalism relates to work on symmetries and equivariance in deep learning [Bronstein et al., 2021], though our focus is on emergent (not engineered) structure.

The gauge-invariance analysis addresses a widespread issue in the representations literature: many reported geometric features of neural network activations depend on the (arbitrary) choice of basis and are therefore not intrinsic properties of the network's computation.

The diagnostic toolkit also connects to causal intervention methods in interpretability [Geiger et al., 2021, Meng et al., 2022], particularly the scrubbing test (Criterion 4), which is a targeted activation intervention.

6.4 Limitations

- All experiments use a single model (Phi-2, 2.7B). Results may not generalize.
- The Bank Experiment tests a specific form of path dependence (topic ordering). Other forms (e.g., reasoning chains, persona priming) are untested.
- Cosine distance and linear probes are Tier 3 diagnostics. The full Tier 2 analysis (whitened pullback metric at intermediate layers) was not performed due to computational cost.
- The linear probe's perfect accuracy with $n = 16$ pairs in $d = 2560$ dimensions should be interpreted cautiously; the feature space vastly exceeds the sample size, and regularization prevents overfitting but does not guarantee the separating hyperplane is meaningful.
- A planned experiment on reflexivity (self-world coupling) was not executed; it remains a direction for future work.

6.5 Broader Implications

The fiber bundle formalism was originally motivated by broader questions about the relationship between internal structure and functional behavior in information-processing systems. We have restricted this paper to representational dynamics in transformer models. The diagnostic toolkit measures a specific, well-defined property (path-dependent internal state variation not captured by output distributions) that is of independent interest for interpretability and alignment research, regardless of any philosophical interpretation. Any broader implications require additional assumptions not addressed here.

7 Conclusion

We have presented a geometric framework for detecting hidden representational structure in transformers, formalized as fiber bundle holonomy. The framework yields gauge-invariant observables, a three-tier classification of computable quantities, and a four-criterion diagnostic test. Applied to Phi-2 (2.7B), all three experiments return negative results: the model does not exhibit the lossy projection from internal states to output distributions that nontrivial fiber structure requires.

The primary contribution is the diagnostic toolkit itself. It defines precisely what "hidden representational structure" means, how to test for it, what controls are needed, and what a positive result would look like. The criteria are model-agnostic and can be applied to any autoregressive transformer.

The four criteria are:

1. Output distributions must match (FR gate).
2. Intermediate-layer divergence must exceed output-layer divergence (non-monotonic profile).
3. Linear probe accuracy must drop from intermediate to output layers (information compression, not rotation).
4. Removing the path-separating direction must not affect output distributions (null-space storage).

Phi-2 fails criteria 1, 3, and 4. We provide this diagnostic toolkit so that as models scale, the emergence of hidden internal structure—should it appear—can be detected rigorously, distinguishing true representational holonomy from spectral artifacts and recency effects.

Acknowledgments

The author thanks Claude Opus 4.6 (Anthropic) and ChatGPT 5.2 (OpenAI) for substantial contributions to mathematical development, experimental design, control analysis, and manuscript preparation.

A Implementation Details

Hardware: Single NVIDIA RTX 3080 (10GB VRAM). All experiments run in half-precision (float16) with `device_map="auto"`.

Software: Python 3.10, `transformers==4.36.2`, `torch==2.1.2` (CUDA 12.1), `scikit-learn==1.3.2`, `numpy==1.26.2`.

Model: Phi-2 (2.7B parameters, 32 layers, $d = 2560$, $V = 51200$, max position embeddings = 2048). Loaded via HuggingFace `transformers` with `trust_remote_code=True`.

Experiment 0: Top-$K = 512$ approximation for the pullback metric $G(z)$. 60 prompts across 9 categories. Row-permutation control uses a fixed random seed (42).

Experiment 1 (Bank): 16 path pairs (4 financial × 4 nature paragraphs). Formatting washout ≈130 tokens. FR gate thresholds: $\{0.02, 0.05, 0.1\}$ with bottom-25% fallback. 30 probes across 4 categories measured via KV-cache reuse. Sense-check probe tokens validated as single-token BPE encodings. Bootstrap confidence intervals: 1000 resamples, seed 42.

Experiment 2 (Layers): All 16 pairs forwarded with `output_hidden_states=True`. Activations extracted at the final token position of the target string. Linear probe: `sklearn.linear_model.LogisticRegr` $C \in \{0.01, 0.1\}$, `solver="lbfgs"`, `max_iter=1000`. Leave-one-group-out cross-validation (groups = pairs). Features standardized per fold. LM-head scrubbing: path direction $\hat{w}$ computed as the normalized mean-difference direction $w = \bar{z}_A - \bar{z}_B$, $\hat{w} = w/\|w\|$ at layer 32; scrubbed activation

$z' = z - (z^\top \hat{w})\hat{w}$; output distribution recomputed via `lm_head`. Recency control: 200-token neutral suffix (technical/data-serialization text, verbatim in repository) appended between washout and target.

Code: Available at https://github.com/anderswrk/representational-holonomy.

References

Shun-ichi Amari. *Information Geometry and Its Applications.* Springer, 2016.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.

Nikolai Nikolaevich Čencov. *Statistical Decision Rules and Optimal Inference.* American Mathematical Society, 1982. Translated from the Russian.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3): 326–334, 1965.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, et al. A mathematical framework for transformer circuits. Transformer Circuits (web publication), 2021. URL https://transformer-circuits.pub/2021/framework/index.html. Accessed 2026-02-07.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12216–12235. Association for Computational Linguistics, 2023.

Mojan Javaheripi and Sébastien Bubeck. Phi-2: The surprising power of small language models. Microsoft Research Blog, 2023. Accessed 2026-02-07.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, 2019.

Jacob A. Zavatone-Veth, Sheng Yang, Julian A. Rubinfien, and Cengiz Pehlevan. How does training shape the Riemannian geometry of neural network representations?, 2023.