HOW TO TRACK YOUR TANGENT SPACE: THE GEOMETRY OF THE LINEAR REPRESENTATION HYPOTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

We examine how local geometry shapes interpretability by studying the tangent structure of the activation manifold of language models: we show that hidden states can be decomposed into linear and multidimensional components by projection onto the tangent space. Using this decomposition, we then show that linear components dominate the interpretable signature of the network. We then perform a similar decomposition of the recently proposed J-lens, to concentrate its on-manifold versus off-manifold data. We give a full analytic description of a tangent-native J-lens, and show that the normal contribution to the J-lens depends on the off-manifold extension of the network: in particular, the workspace readout can be changed arbitrarily subject to mild technical constraints, which could prevent a misaligned network from being detected. Empirically, we show that there is not a simple fix for this: the on-manifold Jacobian shows substantial cancellation in mid-to-late layers, and the J-lens outputs materially rely on their off-manifold projection.

1 INTRODUCTION

Monitoring and safeguarding modern language models relies on the development of powerful interpretability tools, designed to give us an understanding of both the structural mechanisms and the underlying thought processes of these models. The latter has been the subject of a wide array of developments in recent years, including sparse autoencoders (SAEs) (Huben et al., 2024), logit lenses (nostalgebraist, 2020), and linear classifiers. All of these approaches share a basic underlying structure, in which a linear map transforms hidden state values into a more interpretable structure, differing primarily in their training signal.

Underlying these linear decompositions is the *linear representation hypothesis* (Elhage et al., 2022). The linear representation hypothesis posits that network latents can be understood as a sparse linear combination of an underlying representative dictionary. This implies a specific geometric structure of activation space, as a conic, stratified manifold. We observe that there is a simple test for this underlying geometric claim: any latent $h_\ell \in \mathcal{M}_\ell$ will lie in its own tangent space $T_{\mathbf{h}_\ell}\mathcal{M}_\ell$. This is not the case for generic manifolds. In fact, we show that the projection $P_{T_{\mathbf{h}_\ell}\mathcal{M}_\ell}\mathbf{h}_\ell$ decomposes $\mathbf{h}_\ell$ into the approximately linear component of $\mathbf{h}_\ell$ and its nonlinear data. We then show across a synthetic evaluation suite from (Gurnee et al., 2026) that the interpretable content of the latent seems to be concentrated on the tangent space projection across model families.

To evaluate this, we use a recent improvement to the classical logit lens (nostalgebraist, 2020), Gurnee et al. (2026)'s J-lens. The J-lens observes that though the unembedding map U is a powerful decoder on late model layers, it is not well adapted to early-to-mid layer structures in the model. In order to align the spaces, the J-lens calculates the expected Jacobian of the map from layer ℓ to the final layer L of the model, as a strong linear approximation to the downstream causal effect of layers $\ell + 1$ through L. The composite map $U \circ \mathbb{E}[J_{\ell \to L}]$ then gives layer-adapted token probabilities, reflecting the downstream aggregate impact of a hidden state h_ℓ.

The Jacobian in (Gurnee et al., 2026) is defined on the ambient space $\mathbb{R}^d$, but the activation space of the network is much smaller (Ansuini et al., 2019; Valeriani et al., 2023). When localized to the activation manifold, the Jacobian of a map of smooth manifolds $M \to N$ is a bundle map

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

$TM \to TN$, and defines a natural on-manifold J-lens by projecting away normal components of the Jacobian.

This is much easier to describe mathematically than the J-lens of Gurnee et al. (2026): in Theorem 2, we show that when the activation manifold is defined by the linear representation hypothesis, the tangent-projected J-lens is a linear map which sends each primitive of the space to the weighted average of outputs when the primitive is present, minus the average output when the primitive is absent. Importantly, we then show that the full J-lens is not determined by the network's activations on the data manifold. Unlike SAEs, linear probes, or the tangent-restricted J-lens, the full J-lens can be arbitrarily modified along directions not everywhere-tangent to the activation manifold without changing the network's underlying empirical distribution (Theorem 3).

However, our initial experiments showed that the manifold-native J-lens performed substantially worse than a variety of controls: empirically, sufficient cancellation takes place to remove most of the signal from the tangent Jacobian in mid-to-late layers, suggesting that the success reported in Gurnee et al. (2026) depends materially on the specific off-manifold extension implemented by the transformer.

In summary, our contributions are:

- We introduce tangent projection as an intrinsic, local, geometric tool for decomposing language model latents into primarily linear and primarily non-linear components. We then show that linear components of this decomposition control the interpretable directions in the network: **directions which are linearly extensible *and* on-manifold contain most of the interpretable signature measured by the J-lens.**
- We provide a theoretical analysis of the J-lens, showing that it decomposes into a term that depends on the extension of the model in off-manifold directions, and a term that depends on its on-manifold values and the curvature of the activation manifold. **We show that the J-lens can be freely modified, subject to hypotheses, without changing the network's output distribution by perturbing its smooth off-manifold extension.**
- We then show empirically that the performance of the J-lens is driven by its normal extension. In particular, **the tangent projected constructions of the J-lens we test substantially reduce performance**.

2 Related work

Activation manifolds. Many past authors have noted the low dimensional nature of activation manifolds (Ansuini et al., 2019; Valeriani et al., 2023), and the characteristic dimension shifts between network layers. Wurgaft et al. (2026) shows that steering network activations along these manifolds smoothly interpolates model choices, while off-manifold steering can destroy model fluency. Similarly to our analysis of the linear representation hypothesis, Robinson et al. (2024) show that the token space of language models has the structure of a stratified manifold. More recently, authors have explored nonlinear components of the activation manifold (Engels et al., 2025), and proposed mixed decompositions, such as Minkowski sums of low-dimensional manifolds (Bhalla et al., 2026a).

Geometry in interpretability. Zaher et al. (2026) shows that SAE reconstruction quality depends on a number of intrinsic geometric factors, including curvature. Fel et al. (2026) modifies the sparsity structure of SAEs to allow them to capture multidimensional features in natural ways. Sequence-aware SAEs (Lubana et al., 2026; Bhalla et al., 2026b) implicitly approximate tangent and normal components by projecting onto the span of preceding tokens, and passing only the normal component of the hidden state to the SAE.

Tangent & normal constructions. Gradient attributions have previously been observed to be strongest when manifold aligned (Bordt et al., 2023). Most similar to us, Anders et al. (2020) shows that gradient attribution on classifiers can be manipulated by modifying the embedding of the normal space of the data manifold—this is very similar to our Theorem 3, which applies a very similar argument for activation manifolds.

3 BACKGROUND

3.1 NOTATION

We evaluate transformer architecture language models (Vaswani et al., 2017), which alternate between applying attention and perceptron layers. The transformer combines the outputs in a residual stream (Elhage et al., 2021). The residual stream is a sequence of vector spaces $H_\ell \cong \mathbb{R}^d$, where $\ell \in \{0, 1, ..., L\}$ denotes the layer index. The hidden state of a transformer at layer ℓ and token t is a vector $\mathbf{h}_{\ell,t} \in H_\ell$. When t can be inferred from context, we simply write $\mathbf{h}_\ell \in H_\ell$. Given a subspace $V \subset H_\ell$, we write P_V for the orthogonal projection onto V. We write $T_x\mathcal{M}$ for the tangent space at x in $\mathcal{M}$. When it is clear from context, we will abbreviate this by T_x. We will also often write P_T to refer to the projection $P_{T_x\mathcal{M}}$, and similarly for the normal bundle $N\mathcal{M}$.

3.2 THE J-LENS

The logit lens was one of the first language-model specific interpretability tools introduced. It functions by applying the unembedding map of the network U early: the idea is that the network needs to decode its representations into the basis of U at the final layer, and the easiest way to do this is to organize concepts in that basis throughout processing—a similar idea motivates cross-layer transcoders (Lindsey et al., 2024). Gurnee et al. (2026) proposed using the average Jacobian $\mathbb{E}[\frac{\partial \mathbf{h}_{L,t+k}}{\partial \mathbf{h}_{\ell,t}}]$ as a first order approximation to the map of hidden spaces $H_\ell \to H_L$, emphasizing causality of the hidden states on the outputs. The Jacobian naturally decomposes into on-manifold directions and normal directions: we explore how this local geometry shapes the J-lens in Section 5.

3.3 TANGENT SPACES

Given a smooth embedding $\iota : \mathcal{M} \to \mathbb{R}^d$ of a k dimensional manifold into Euclidean space, the tangent space at $x \in \mathcal{M}$ is the k dimensional vector space $T_x\mathcal{M}$ which is parallel to the manifold at x. This is the natural place where derivatives operate, providing an underlying linear structure at each point of the manifold. The tangent bundle $T\mathcal{M}$ is the union of the tangent spaces at all $x \in \mathcal{M}$. The normal space $N_x\mathcal{M}$ at $x \in \mathcal{M}$ is the $d-k$ dimensional vector space orthogonal to $T_x\mathcal{M}$, going through x. This measures perturbations in the ambient space $\mathbb{R}^d$ that leave the manifold.

4 THE GEOMETRY OF THE LINEAR REPRESENTATION HYPOTHESIS

The linear representation hypothesis is one of the most enduring ideas in interpretability (Elhage et al., 2022; Park et al., 2023; Gurnee & Tegmark, 2024). In this model, hidden states decompose into sparse sums of primitive concept vectors,

$$\mathbf{h} = \sum_{j \in J \subset I,\ |J| \leq k \ll |I|} a_j \mathbf{c}_j \qquad a_j > 0,\ \mathbf{c}_i \cdot \mathbf{c}_j \approx \delta_{ij}. \tag{1}$$

For example, the word "Paris" could be represented as a sum of concepts including "France," "capital cities," and "art," but is unlikely to incorporate concepts like "baseball." If the number of concepts $|I|$ is less than the ambient dimension d, then each concept can be orthogonally embedded, resulting in zero interference when a relevant concept needs to be retrieved. In language models, however, we generally hypothesize that the number of atomic concepts $\mathbf{c}_i$ greatly exceeds the representation dimension of the model, so we can demand only $\mathbf{c}_i \cdot \mathbf{c}_j \approx \delta_{ij}$.

Geometrically, equation (1) describes a *conic stratified manifold*: it is a union of k-dimensional rays from the origin, which meet one another (non-smoothly) at $(k-1)$-dimensional rays, their strata, which in turn intersect in $(k-2)$-dimensional rays, and so on. For $k = 2$ and $d = 3$, this would resemble honeycomb.

Recently, several authors have studied multidimensional (Fel et al., 2026) and nonlinear (Engels et al., 2025) components of feature space. If, in equation (1), some concepts have heavily correlated coefficients across the data-space, they can create non-linear structures (Bhalla et al., 2026a). In our honeycomb analogy, a heavy (anti-)correlation between the coefficients on $\mathbf{c}_i$ and $\mathbf{c}_j$ would turn the

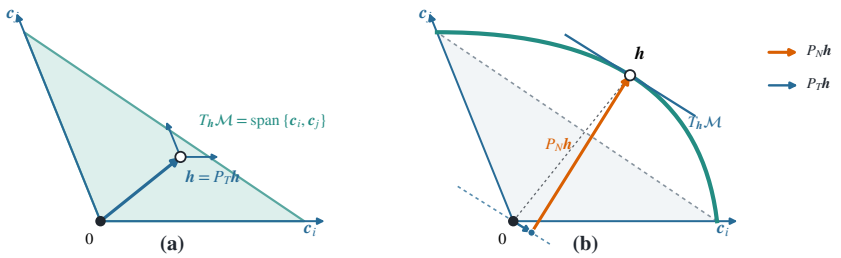

Figure 1: (a) Linearly represented concepts exist on the tangent plane to the manifold. (b) Nonlinear components don't project cleanly onto the tangent space, but incorporate substantial normal data.

wall between vertices i and j into a hoop. In this form, which we call the *weak linear representation hypothesis*, the activation manifold can be described instead as the set of points

$$\mathbf{h} = \underbrace{\sum_{j \in J \subset I,\ |J| \leq k \ll |I|} a_j \mathbf{c}_j}_{\text{LRH components}} + \underbrace{\sum_{r \in R \subset S, |R| \leq q \ll |I|+|S|} \mathbf{p}_r}_{\text{Multidimensional terms}} \qquad a_j > 0,\ \mathbf{c}_i \cdot \mathbf{c}_j \approx \delta_{ij}, \mathbf{p}_r \in r, \tag{2}$$

or a *Minkowski sum* of nearly-orthogonal manifolds, only some of which are rays.

Since the $\mathbf{c}_i, \mathbf{p}_r$ are approximately orthogonal to one another, it's very simple to describe the local geometry around a generic point on this stratified manifold $\mathbf{h} \in \mathcal{M}$: the tangent space $T_\mathbf{h}\mathcal{M}$ has an almost-orthogonal basis given by the linear components $\{\mathbf{c}_j \mid j \in J\}$ and the tangent spaces to the multidimensional components $\{T_{\mathbf{p}_r} r \mid r \in R\}$. Importantly, if the muldimensional component doesn't locally break down as a free sum of linear components, these latter terms are structurally different than the former terms, in that $\mathbf{c}_j \in \text{Span}(\mathbf{c}_j)$, but $\mathbf{p}_r \notin T_{\mathbf{p}_r} r$. In simple cases like the sphere, we in fact have $P_{T_{\mathbf{p}_r} r}\mathbf{p}_r = 0$, and generically the projection of $\mathbf{p}_r$ onto its own tangent space decreases the norm. The reverse is true for projection onto the normal space: for linear components: $P_{N_{\mathbf{c}_j}\mathcal{M}}\mathbf{c}_j = 0$.

As a result, even without access to the true decomposition (2) of $\mathcal{M}$, we can nevertheless approximate the decomposition of a vector $\mathbf{h}$ into linear and multidimensional terms by calculating its projection into $T_\mathbf{h}\mathcal{M}$ and $N_\mathbf{h}\mathcal{M}$.

Although the projection depends on the choice of origin and ambient metric, these are not arbitrary: the use of RMSNorm and additivity of the residual stream privileges the 0-vector and equips readouts with Euclidean-normalized states. Some work has been performed on alternative intrinsic metrics (Park et al., 2023) for interior layers, which is an interesting avenue for future work.

Remark 1. Note this is only an approximation: without imposing stronger assumptions on the nonlinear manifolds that appear, we cannot guarantee that they will be captured by normal projection. Instead, this is a heuristic measure designed to partially isolate the LRH components of the decomposition.

4.1 Tangent spaces and curvature

The activation manifold $\mathcal{M}_\ell$ is the subset of the hidden space H_ℓ in which activations empirically occur. We study two methods of approximating the tangent space to $\mathbf{h}_\ell$ in $\mathcal{M}_\ell$:

- **Pushforwards:** The secant space in the embedding space H_0 is easy to approximate: H_0 contains a finite, discrete set of tokens. We use PCA on centered sets of k-nearest neighbors to make approximations of the secant space of the embedding cloud in H_0. We then calculate the tangent space in layer ℓ as the pushforward of this bundle in H_0—these are precisely the *upstream* Jacobians $\frac{\partial \mathbf{h}_{\ell,t}}{\partial \mathbf{h}_{0,t-j}}$, projected onto the tangent space of H_0. We calculate the tangent space by applying SVD to these matrices, and preserving a fraction p of the energy (Shao et al., 2018).

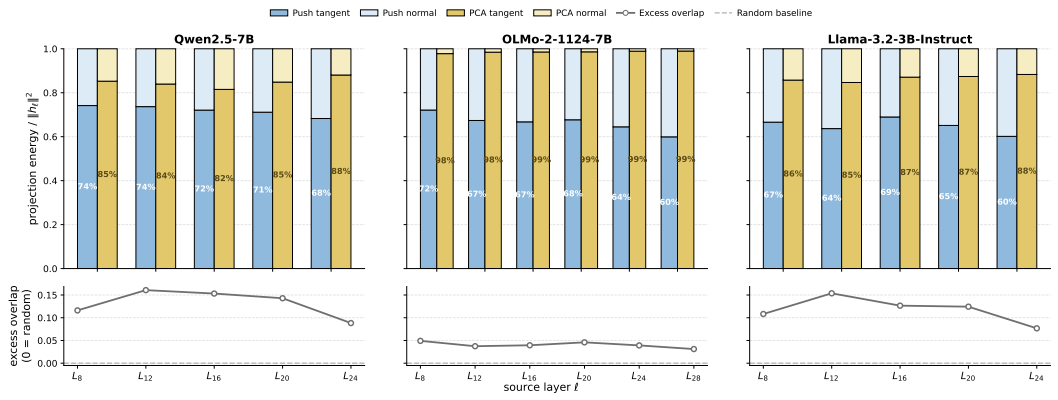

Figure 2: Projection energy $\frac{||P_{T_{\mathbf{h}_\ell}\mathcal{M}}\mathbf{h}_\ell||^2}{||\mathbf{h}_\ell||^2}$ of $\mathbf{h}_\ell$ in the approximated tangent space $T_{h_\ell}\mathcal{M}$ for Qwen-2.5 7B, OLMo-2 7B, and LLaMA-3 3B. Below, we show the excess overlap between the pushforward tangent space and the PCA tangent space.

- **PCA:** We take the k nearest neighbors of $\mathbf{h}_\ell$ from a precomputed corpus, and perform PCA directly in H_ℓ, preserving a fraction p of the energy of the system (Tyagi et al., 2013).

Both of these are good notions of tangent space, but they measure moderately different things. The pushforward tests how hypothetical variations in the prompt space would perturb the embedding, while direct PCA works in realized activations. This latter point is important for emergent geometric forms in language models, such as circles representing days of the week (and more broadly, and cyclical time structure) and hues (Bhalla et al., 2026a; Engels et al., 2025), which manifest only in mid-network layers. The downside to the second approach is that its accuracy depends on the creation of a sufficiently large corpus, which in high dimensions can be prohibitive, while the pushforward can be calculated purely from the input and the parameters of the network.

4.2 Empirical decompositions

Using the decompositions of Section 4.1, we tested Qwen-2.5 7B (Team et al., 2025), OLMo-2 7B (OLMo et al., 2024), and LLaMA-3.2 3B (Dubey et al., 2024), and summarize results in Figure 2. Across layers in all evaluated models we find that PCA captures more of the variance of $\mathbf{h}_\ell$ than the pushforward does. This is particularly true for OLMo-2, which captures $\sim 98\%$ of the variance of h_ℓ in its tangent space. The two approximations also capture quite different spaces, which we measured by excess overlap (0 = random subspaces, 1 = one space is a subspace of the other). While Qwen and LLaMA show noteworthy overlap, OLMo-2 maintains an excess overlap of only ≈ 0.05.

Across all models, the projection energy is much higher than would be expected by chance, suggesting a strong linear component in the decomposition (2).

4.3 Useful data is highly concentrated along the tangent space

To measure the semantic split between the linear components of $\mathbf{h}$ and the nonlinear components, we used the J-lens (Gurnee et al., 2026) as a training-free interpretability tool, applied to the projections $P_{T_{\mathbf{h}_\ell}\mathcal{M}}\mathbf{h}_\ell$ and $P_{N_{\mathbf{h}_\ell}\mathcal{M}}\mathbf{h}_\ell$, and monitored results on six synthetic readout suites (see Appendix B for details).

Results for the pushforward tangent space are summarized in Table 1, and for PCA tangent spaces in Table 2: across all three model families, the pushforward tangent projection has a better median rank score than the full Jacobian map, and similarly for the PCA tangent projection with the exception of LLaMA. Meanwhile the normal bundle projection collapses results. Pass rates are similar for the tangent projection and the full vector $\mathbf{h}_\ell$, while substantially worse for the normal projection. On Qwen, the tangent projection has a denoising effect: $\mathbb{E}[J]P_T(\mathbf{h}_\ell)$ is closer to the true map induced

Table 1: Pass rates and median rank on the synthetic lens evaluation suite of (Gurnee et al., 2026), applied to the baseline hidden state and its (pushforward) tangent and normal projections, averaged over all suites. These evaluations measure the presence of intermediate and related concepts in the hidden state. Particularly in Qwen and LLaMA, this semantic data is heavily concentrated in the tangent projection $P_T(\mathbf{h}_\ell)$.

Model	Projection	P@1 ↑	P@5 ↑	P@10 ↑	Mean med. rank ↓	KL ↓	Cos. ↑
Qwen-2.5 7B	$\mathbf{h}_\ell$	0.195	0.356	0.407	461.8	10.429	0.405
	$P_T(\mathbf{h}_\ell)$	**0.267**	**0.437**	**0.546**	**137.0**	**9.286**	**0.470**
	$P_N(\mathbf{h}_\ell)$	0.002	0.011	0.011	9609.4	15.801	-0.098
OLMo-2 7B	$\mathbf{h}_\ell$	**0.204**	**0.395**	**0.507**	251.3	**7.299**	**0.485**
	$P_T(\mathbf{h}_\ell)$	0.187	0.363	0.419	**169.5**	7.801	0.422
	$P_N(\mathbf{h}_\ell)$	0.142	0.277	0.352	528.1	7.899	0.366
LLaMA-3.2 3B	$\mathbf{h}_\ell$	0.305	**0.480**	**0.604**	23.9	**4.719**	**0.606**
	$P_T(\mathbf{h}_\ell)$	**0.320**	0.473	0.571	**19.4**	4.842	0.573
	$P_N(\mathbf{h}_\ell)$	0.065	0.134	0.208	278.5	7.453	0.263

Table 2: Held-out evaluation results for Qwen, OLMo, and LLaMA with direct PCA tangent spaces. Pass@k and median rank are computed with the highest activating passing token, with Pass@k aggregated over all sequences, and median rank averaged across the six suites.

Model	Projection	P@1 ↑	P@5 ↑	P@10 ↑	Mean med. rank ↓	KL ↓	Cos. ↑
Qwen-2.5 7B	$\mathbf{h}_\ell$	0.195	0.356	0.407	461.8	10.429	0.405
	$P_T(\mathbf{h}_\ell)$	**0.205**	**0.379**	**0.439**	**193**	**10.264**	**0.437**
	$P_N(\mathbf{h}_\ell)$	0.000	0.000	0.000	14437.6	17.148	-0.053
OLMo-2 7B	$\mathbf{h}_\ell$	**0.204**	**0.395**	**0.507**	251.3	**7.299**	**0.485**
	$P_T(\mathbf{h}_\ell)$	0.199	0.372	0.496	**231.6**	7.385	0.483
	$P_N(\mathbf{h}_\ell)$	0.000	0.000	0.000	3532.3	9.374	0.173
LLaMA-3.2 3B	$\mathbf{h}_\ell$	**0.305**	**0.480**	**0.604**	**23.9**	**4.719**	**0.606**
	$P_T(\mathbf{h}_\ell)$	0.301	0.447	0.519	28.7	5.299	0.576
	$P_N(\mathbf{h}_\ell)$	0.145	0.222	0.233	576.3	7.514	0.290

by the transformer than the unprojected variant $\mathbb{E}[J]\mathbf{h}_\ell$ in terms of both KL divergence and cosine similarity.

The performance of this decomposition on OLMo is notably worse than Qwen and LLaMA. Given the results of Figure 2, where the excess overlap with the PCA tangent space is quite low, we hypothesize that our pushforward tangent approximation is poor on OLMo-2. OLMo obtains more similar results to Qwen and LLaMA with the PCA tangent space. These results, along with breakdowns by evaluation suite, can be found in Appendix C.

5 Decomposing the Jacobian lens along the tangent space

Given the success of the local projection $\mathbb{E}[J]P_{T_{\mathbf{h}_\ell}}\mathbf{h}_\ell$ in Section 4.3, we hypothesized that by commuting the expectation to study $\mathbb{E}[JP_{T_{\mathbf{h}_\ell}}]$, we could construct a Jacobian lens that acts purely on-manifold. Mathematically, this approach is well founded: the Jacobian $\frac{\partial h_{L,t+k}}{\partial h_{\ell,t}}$ is a map $T\mathcal{M}_\ell \to T\mathcal{M}_L$ from the tangent bundle on the input space to the tangent bundle on the output space. For each point $\mathbf{h}_\ell \in \mathcal{M}_\ell$, this describes how local directions on $\mathcal{M}_\ell$ translate to directions on $\mathcal{M}_L$. Several authors have noted that on-manifold translations better preserve model capacity (Wurgaft et al., 2026), so this projection may be more intrinsic.

When the Jacobian lens

$$\mathbb{E}[J_{\ell\to L}] = \mathbb{E}_{\mathbf{h}_\ell,k}\left[\frac{\partial \mathbf{h}_{L,t+k}}{\partial \mathbf{h}_{\ell,t}}\right] = \frac{1}{n}\sum_{k=0}^{n-1}\int_{\mathcal{M}_\ell}\frac{\partial \mathbf{h}_{L,t+k}}{\partial \mathbf{h}_{\ell,t}}d\mu \tag{3}$$

is calculated, we don't have direct access to the activation manifold $\mathcal{M}_\ell$, so the Jacobians are calculated as a map $\mathbb{R}^d \to \mathbb{R}^d$. The contribution from each datum is therefore $\frac{\partial \mathbf{h}_{L,t+k}}{\partial \mathbf{h}_{\ell,t}} = \frac{\partial \mathbf{h}_{L,t+k}}{\partial \mathbf{h}_{\ell,t}}|_{T_{\mathbf{h}_\ell}\mathcal{M}} + \frac{\partial \mathbf{h}_{L,t+k}}{\partial \mathbf{h}_{\ell,t}}|_{N_{\mathbf{h}_\ell}\mathcal{M}}$, and we can decompose the full J-lens as

$$\int_{\mathcal{M}_\ell} \frac{\partial \mathbf{h}_{L,t+k}}{\partial \mathbf{h}_{\ell,t}} d\mu = \int_{\mathcal{M}_\ell} \frac{\partial \mathbf{h}_{L,t+k}}{\partial \mathbf{h}_{\ell,t}} P_{T_{\mathbf{h}_\ell}\mathcal{M}} d\mu + \int_{\mathcal{M}_\ell} \frac{\partial \mathbf{h}_{L,t+k}}{\partial \mathbf{h}_{\ell,t}} P_{N_{h_\ell}\mathcal{M}} d\mu. \tag{4}$$

The tangent projected term in Equation (4) can be computed using Stokes' Theorem. If $\mathcal{M}_\ell$ is a smooth manifold, we can calculate the contribution as

Theorem 1. *Suppose $\mathcal{M}_\ell$ is a smooth manifold of positive codimension with boundary $\partial\mathcal{M}_\ell$, and $d\mu = \rho\, dV_{\mathcal{M}}$ is a smooth sampling density on M_ℓ. Let $f : \mathcal{M}_\ell \to \mathcal{M}_L$ denote the induced map of manifolds with integrable derivative over $\mathcal{M}$, and suppose either $\mathcal{M}$ is compact or $f\rho$ has sufficient decay at infinity to be approximated by compact subregions. Then*

$$\int_{\mathcal{M}_\ell} \frac{\partial \mathbf{h}_{L,t+k}}{\partial \mathbf{h}_{\ell,t}} P_{T_{\mathbf{h}_\ell}\mathcal{M}} d\mu = \int_{\partial\mathcal{M}} f(\mathbf{h}_\ell)\rho(\mathbf{h}_\ell)\nu(\mathbf{h}_\ell)^\top dS - \int_{\mathcal{M}_\ell} f(\mathbf{h}_\ell)\mathcal{H}^\top d\mu, \tag{5}$$

where $\mathcal{H}$ is the weighted mean-curvature $\mathcal{H} = \Delta_{\mathcal{M}_\ell}\iota + \operatorname{grad}_{\mathcal{M}} \log \rho$ of the inclusion $\iota : \mathcal{M}_\ell \to \mathbb{R}^d$, with $\Delta_{\mathcal{M}_\ell}$ the Laplacian, and ν is the outward conormal vector along $\partial\mathcal{M}_\ell$.

For the conic stratified manifold defined by the (strong) linear representation hypothesis (equation (1)), we can describe the Jacobian explicitly:

Theorem 2. *Suppose $C_J = \{\sum_{j\in J} a_j \mathbf{c}_j \mid a_j \geq 0\}$ is a top-dimensional cone in the linear representation manifold, with $\mathbf{c}_j$ orthonormal, and μ is a restricted half-normal distribution with $\rho(\sum a_i \mathbf{c}_i) = \prod_i h_{\sigma_i}(a_i)$. Then the tangent projected Jacobian from C_J is*

$$\sum_i \frac{1}{\sigma_i^2} \mathbb{E}_{C_J}\left[a_i \left(f\left(\sum_j a_j \mathbf{c}_j\right) - f\left(\sum_{j\neq i} a_j \mathbf{c}_j\right)\right)\right] \mathbf{c}_i^\top. \tag{6}$$

Intuitively, this is the rank (at most) $|J|$ linear map which takes $\mathbf{c}_i$ to the weighted average of its outputs, minus the weighted average on the cone when $\mathbf{c}_i$ is not present.

While the details for the tangent component resolve very nicely, the second term of Equation (4) depends on the extension of the model off the activation manifold. In particular, the normal component can differ for two models with identical in-distribution empirical distributions, and we can perturb the normal directions without changing the sampling distribution.

Theorem 3. *Let $\mathcal{M} \subset \mathbb{R}^d$ be a properly embedded manifold of positive codimension, μ a probability measure on $\mathcal{M}$, and $f : \mathbb{R}^d \to \mathbb{R}^m$ smooth with integrable derivative. Let $Q = \int_{\mathcal{M}} P_{N_x\mathcal{M}}\, d\mu$, and $K = \ker Q$. Then for any linear map $B : \mathbb{R}^d \to \mathbb{R}^m$ such that $BK = 0$, there is a smooth map $g : \mathbb{R}^d \to \mathbb{R}^m$ such that $g|_{\mathcal{M}} = f|_{\mathcal{M}}$, and $\int_{\mathcal{M}} DgP_N\, d\mu = B + \int_{\mathcal{M}} Df P_N\, d\mu$. In particular, the Jacobian lens can be freely modified by any matrix B that annihilates $K = \{v \in \mathbb{R}^d \mid v \in T_x\mathcal{M}$ almost everywhere$\}$.*

We are not the first to observe that gradient methods are susceptible to normal-bundle based attacks: it was originally observed by Anders et al. (2020), who proved the statement:

Theorem 4 (Anders et al. (2020))**.** *Let $\mathcal{M} \subset \mathbb{R}^d$ be a smoothly embedded manifold with positive codimension, and let $\{V_i\}_{1\leq i\leq m}$ be conservative vector fields in $N\mathcal{M}$. Then there is a map $r : \mathbb{R}^d \to \mathbb{R}^m$ such that $r|_{\mathcal{M}} = 0$ and the gradient of r on $N\mathcal{M}$ is given coordinatewise by the V_i on $N\mathcal{M}$.*

These are variants of the submanifold extension lemma, which is well known in differential geometry. These results also only guarantee the existence of such perturbations among smooth functions; we leave the question of realization in standard weight spaces to future work.

In practice, implicit and explicit regularization in language models constrain the realized normal data, and our experiments below find that the normal component is highly informative. This means the standard J-lens of Gurnee et al. (2026) can potentially be manipulated without changing the

Table 3: Aggregate held-out performance of the full, tangent, and normal Jacobian lenses. Pass@k and median rank use the best target-token rank over workspace layers, Pass@k is pooled over all data, median rank is averaged over the six suites; KL is averaged over workspace layers and then examples. See Appendix B for details. Note that results for $\mathbb{E}[J]$ differ slightly here from Table 1 as the training dataset for the projected Jacobians was smaller than what was used for local experiments.

Model	Projection	P@1 ↑	P@5 ↑	P@10 ↑	KL ↓	Mean med. rank ↓
Qwen-2.5 7B	$\mathbb{E}[J]$	**0.205**	**0.369**	**0.404**	**10.00**	393.3
	$\mathbb{E}[JP_T]$	0.127	0.221	0.272	11.10	**309.5**
	$\mathbb{E}[JP_N]$	0.146	0.286	0.348	10.66	674.8
OLMo-2 7B	$\mathbb{E}[J]$	0.224	0.391	**0.477**	**7.96**	266.8
	$\mathbb{E}[JP_T]$	0.062	0.146	0.189	7.99	948.3
	$\mathbb{E}[JP_N]$	**0.237**	**0.404**	0.474	8.46	**181.6**
LLaMA-3.2 3B	$\mathbb{E}[J]$	0.337	**0.509**	**0.596**	**4.76**	**20.6**
	$\mathbb{E}[JP_T]$	0.143	0.286	0.367	5.23	92.3
	$\mathbb{E}[JP_N]$	**0.350**	0.499	0.566	5.61	32.8

distribution of network outputs. As a monitoring tool for recursively self-improving agents, this is important: in principle, **a misaligned actor could arbitrarily change its J-lens readouts on directions that are not almost everywhere tangent to $\mathcal{M}$ without changing its underlying sampling distribution by modifying the off-manifold signature of the network.**

5.1 The J-lens relies on task-relevant normal components

Given the results of Section 4.1 and Theorems 1–3, we hypothesized that the tangent component of Equation (4) might dominate the signal found by the J-lens. To verify this, we ran a range of experiments with the pushforward tangent space approximation on both $\int_{M_\ell} \frac{\partial h_{L,t+k}}{\partial h_{\ell,t}} \mid_{T_{h_\ell} M} d\mu$ and $\int_{M_\ell} \frac{\partial h_{L,t+k}}{\partial h_{\ell,t}} \mid_{N_{h_\ell} M} d\mu$. Across our studied variations, however, we found the opposite was true: much of the explanatory power of the J-lens comes from the integrated normal directions. We summarize these results in Table 3. Interestingly, while performance on evaluation suites is task-dependent and broadly concentrated on the normal projection, this is not matched in KL divergence of the output as compared to the network, where the tangent projection achieves similar results.

Across most aggregate metrics, the normal-projected lens outperforms the tangent-projected lens, though the two exhibit complementary task- and model- dependent utility. The clearest cross-model exception to the results of Table 3 was the `order-ops` suite, which measures the formation of intermediate values while evaluating expressions. A full decomposition by benchmark is given in Appendix C. Our key claim is therefore that the normal contribution cannot be ablated without damaging the broad performance of the J-lens, not that the tangent-projected lens does not contain information.

To understand why the tangent performance deteriorated, we studied the relative Frobenius norm coherence of the projected operators. While the pointwise Frobenius norm of both the tangent and normal components remain commensurate with the rank of the bundle (see Appendix C), the norm of the average decreases more sharply as layer index increases. We show this in Figure 3: here *coherence* of an embedded vector field v is the ratio $\frac{||\frac{1}{n}\sum v(x_i)||^2}{\frac{1}{n}\sum ||v(x_i)||^2}$. The relative coherence is the ratio between the coherence of the tangent (resp. normal) Jacobian vector field and the coherence of the unprojected Jacobian vector field, using the Frobenius norm.

This trend goes some distance in explaining the contrast between our results and (Anders et al., 2020), which found tangent projection to be robust: they studied tangent projection primarily on the input data manifold, while we study it on intermediate activations. Since coherence deteriorates in late layers, it may not have been a confounding factor for earlier studies. Since most readout results were concentrated in layers 20 and later, the cancellation of certain directions in the hidden space is consistent with the result.

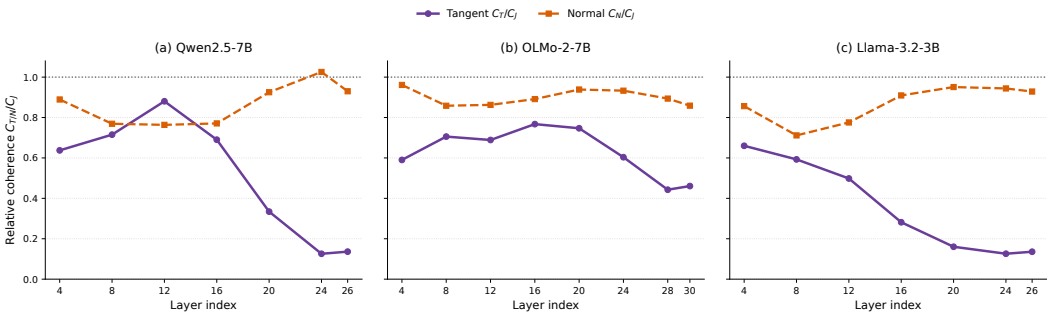

Figure 3: Relative coherence of the tangent and normal projected Jacobians across layers (1=relative coherence of the full Jacobian). As layer index increases, relative coherence of the tangent-projected Jacobian substantially decreases, particularly for Qwen and LLaMA.

6 Discussion

We studied the local implications of the linear representation hypothesis, and its implied conic, stratified geometry. Curved, nonlinear data can be differentiated from linearly encoded data by studying the tangent structure of the activation manifold, which we approximated using both pushforward bundles and in-layer activations. In this decomposition, we found an important distinction: data measured by a battery of interpretability evaluations was primarily concentrated in extensible linear directions: that is, components of a latent that are parallel to the tangent space at that point.

Based on this decomposition, we then studied the decomposition of the J-lens along the tangent and normal spaces. We gave a theoretical description of the on-manifold Jacobians in terms of the linear representation hypothesis, and showed that the utility of the J-lens depends materially on its off-manifold extension. We showed this is driven by late-layer cancellation of tangent Jacobian data. Notably, the off-manifold extension of a network is not determined by its on-manifold data: two networks with identical empirical distributions need not have the same J-lens, and when there is no direction that is tangent to $\mathcal{M}$ almost everywhere, it is possible to arbitrarily modify the J-lens without changing the output distribution of the network under certain hypotheses.

AI use statement

While preparing this work we used generative AI tools for code generation, bug fixes, verifying, correcting, and improving mathematical arguments, including proposing the current form of Theorem 3, assistance in writing proofs, to verify and amend our data interpretations, to search for related work, to create Figure 1, and for final proofreading. We verified all code, and wrote all final arguments, including both mathematical claims and the text of the paper, and we verified the relationship of all related work suggested. We have not used generative AI tools to propose hypotheses or research methodologies, for translation, or for synthetic dataset creation. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This paper discusses methods of evading gradient-based interpretability methods (Theorem 3), which has potential for misuse. We believe that it is nevertheless in the interest of the field to publish these results: our methods are not difficult to re-derive, and it is important for model auditors to be aware of the potential pitfalls in interpretability tools.

Reproducibility statement

To ensure reproducibility of our work, we have attached a code base, including scripts to reproduce all experiments described in the paper. In the non-anonymized release, this is published as a public git repository. We have also fully described hyperparameters of our experiments in Appendix B, and

detailed the precise mathematical interventions we perform, so they can be independently recreated without our repository. Likewise, proofs of all mathematical claims are given in Appendix A.

REFERENCES

Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pp. 314–323. PMLR, 2020.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, Tal Haklay, Daniel Wurgaft, Siddharth Boppana, Matthew Kowal, Vasudev Shyam, Jack Merullo, Atticus Geiger, and Ekdeep Singh Lubana. Do sparse autoencoders capture concept manifolds?, 2026a.

Usha Bhalla, Alex Oesterling, Claudio Mayrink Verdun, Hima Lakkaraju, and Flavio Calmon. Temporal sparse autoencoders: Leveraging the sequential nature of language for interpretability. In *International Conference on Learning Representations*, volume 2026, pp. 141147–141175, 2026b.

Sebastian Bordt, Uddeshya Upadhyay, Zeynep Akata, and Ulrike von Luxburg. The manifold hypothesis for gradient-based explanations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3697–3702. IEEE, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Ashish Kadian, Ahmad Al-Dahle, Aiesha Letkovich, Anirudh Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

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

Josh Engels, Eric Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, volume 2025, pp. 84591–84622, 2025.

Thomas Fel, Matthew Kowal, Mozes Jacobs, Dron Hazra, Usha Bhalla, Lee Sharkey, Lucius Bushnaq, Satchel Grant, Tal Haklay, Thomas Icard, et al. Structuring sparsity: Block-sparse featurizers capture visual concept manifolds. *arXiv preprint arXiv:2606.25234*, 2026.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, volume 2024, pp. 2483–2503, 2024.

Wes Gurnee, Nicholas Sofroniew, et al. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. URL `https://transformer-circuits.pub/2026/workspace/index.html`.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, 25, 2024.

Ekdeep Singh Lubana, Can Rager, Sai Sumedh R Hindupur, Valérie Costa, Oam Patel, Sonia Murthy, Thomas Fel, Greta Tuckute, Daniel Wurgaft, Eric Bigelow, et al. Priors in time: Missing inductive biases for language model interpretability. In *International Conference on Learning Representations*, volume 2026, pp. 42231–42290, 2026.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Michael Robinson, Sourya Dey, and Shauna Sweet. The structure of the token space for large language models. *arXiv preprint arXiv:2410.08993*, 2024.

Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 428–4288. IEEE, 2018.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

Hemant Tyagi, Elıf Vural, and Pascal Frossard. Tangent space estimation for smooth embeddings of riemannian manifolds®. *Information and Inference: A Journal of the IMA*, 2(1):69–114, 2013.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphael Sarfati, Thomas McGrath, et al. Manifold steering reveals the shared geometry of neural network representation and behavior. *arXiv preprint arXiv:2605.05115*, 2026.

Eslam Zaher, Maciej Trzaskowski, Quan Nguyen, and Fred Roosta. The geometric wall: Manifold structure predicts layerwise sparse autoencoder scaling laws. *arXiv preprint arXiv:2605.09887*, 2026.

A Proofs

A.1 Theorem 1

Theorem. *Suppose $\mathcal{M}_\ell$ is a smooth manifold of positive codimension with boundary $\partial\mathcal{M}_\ell$, and $d\mu = \rho\, dV_{\mathcal{M}}$ is a smooth sampling density on M_ℓ. Let $f : \mathcal{M}_\ell \to \mathcal{M}_L$ denote the induced map of manifolds with integrable derivative over $\mathcal{M}$, and suppose either $\mathcal{M}$ is compact or $f\rho$ has sufficient decay at infinity to be approximated by compact subregions. Then*

$$\int_{\mathcal{M}_\ell} \frac{\partial \mathbf{h}_{L,t+k}}{\partial \mathbf{h}_{\ell,t}} P_{T_{\mathbf{h}_\ell}\mathcal{M}} d\mu = \int_{\partial\mathcal{M}} f(\mathbf{h}_\ell)\rho(\mathbf{h}_\ell)\nu(\mathbf{h}_\ell)^\top dS - \int_{\mathcal{M}_\ell} f(\mathbf{h}_\ell)\mathcal{H}^\top d\mu, \tag{7}$$

where $\mathcal{H}$ is the weighted mean-curvature $\mathcal{H} = \Delta_{\mathcal{M}_\ell}\iota + \mathrm{grad}_{\mathcal{M}} \log\rho$ of the inclusion $\iota : \mathcal{M}_\ell \to \mathbb{R}^d$, with $\Delta_{\mathcal{M}_\ell}$ the Laplacian, and ν is the outward conormal vector along $\partial\mathcal{M}_\ell$.

For simplicity, assume f is scalar valued—we can recover the general form by repeating the argument d times. Let $X_j = P_{T_x\mathcal{M}} e_j$ be the jth coordinate vector field on $\mathcal{M}$, and $\eta_j = \iota_{X_j}(\rho dV_{\mathcal{M}})$, where ι denotes interior multiplication, so η_j is a codegree 1 form. Then $df(X_j)\rho dV_{\mathcal{M}} = df \wedge \eta_j$ is the top-degree form we want to integrate, so

$$\int_{\mathcal{M}} (df)_j d\mu = \int_{\mathcal{M}} df \wedge \eta_j. \tag{8}$$

By the product rule, $d(f \wedge \eta_j) = df \wedge \eta_j + f \wedge d\eta_j$, so using Stokes' theorem on the left hand side and rearranging terms, we find

$$\int_{\mathcal{M}} df \wedge \eta_j = \int_{\partial\mathcal{M}} f \wedge \eta_j - \int_{\mathcal{M}} f \wedge d\eta_j. \tag{9}$$

By Cartan's magic formula, $d\eta_j = \operatorname{div}(\rho X_j) dV_{\mathcal{M}} = (\rho \operatorname{div}_{\mathcal{M}} X_j + \langle \operatorname{grad}_{\mathcal{M}} \rho, X_j \rangle) dV_{\mathcal{M}} = \rho(\operatorname{div}_{\mathcal{M}} X_j + \langle \operatorname{grad}_{\mathcal{M}} \log \rho, X_j \rangle) dV_{\mathcal{M}}$. Since $X_j = \operatorname{grad}_{\mathcal{M}}(q_j)$, where q_j is the map $q_j(x) = x_j$, the first term is equal to the Laplacian $\Delta_{\mathcal{M}} q_j =: \mathcal{H}_j$, or the jth component of the mean-curvature vector.

The boundary contribution is concentrated on the outward conormal vectors ν, that is, the unit vectors that lie in $T\mathcal{M}$, but are normal to $T\partial\mathcal{M}$. We've now recovered the claimed equation,

$$\int_{\mathcal{M}} Df\, P_T(\mathbf{h}) d\mu(\mathbf{h}) = \int_{\partial M} f(\mathbf{h})\rho(\mathbf{h})\nu(\mathbf{h})^\top dS - \int_{\mathcal{M}} f(\mathbf{h})\mathcal{H}(\mathbf{h})^\top d\mu(\mathbf{h}). \tag{10}$$

A.2 Theorem 2

Theorem. *Suppose $C_J = \{\sum_{j\in J} a_j \mathbf{c}_j \mid a_j \geq 0\}$ is a top-dimensional cone in the linear representation manifold, with $\mathbf{c}_j$ orthonormal, and μ is a restricted half-normal distribution with $\rho(\sum a_i \mathbf{c}_i) = \prod_i h_{\sigma_i}(a_i)$. Then the tangent projected Jacobian from C_J is*

$$\sum_i \frac{1}{\sigma_i^2} \mathbb{E}_{C_J} \left[a_i \left(f\left(\sum_j a_j \mathbf{c}_j \right) - f\left(\sum_{j\neq i} a_j \mathbf{c}_j \right) \right) \right] \mathbf{c}_i^\top. \tag{11}$$

Though the cone is a manifold with boundaries, Stokes' Theorem does extend to this setting, and corner strata are measure 0 in the boundary measure.

We need to simplify the result of Theorem 1 for the linear representation (stratified) manifold. Since the stratum C_J is flat, the mean-curvature $\Delta_{\mathcal{M}_\ell} \iota = 0$. Therefore the score is $\operatorname{grad}_{\mathcal{M}} \log \mu = \sum_i \frac{\partial}{\partial a_i} \log \mu \mathbf{c}_i^\top = \sum_i -\frac{a_i}{\sigma_i^2} \mathbf{c}_i^\top$. We can therefore write the second term in equation (7) as $\sum \frac{1}{\sigma_i^2} \mathbb{E}[a_i f(\sum a_i \mathbf{c}_i)] \mathbf{c}_i^\top$.

The boundary of C_J is the set of faces $\partial_i C_J$ where $a_i = 0$. The outward unit vector along this boundary is $-\mathbf{c}_i$, so the boundary contribution is $-\mathbb{E}[\mu(\sum a_j \mathbf{c}_j) f(\sum a_j \mathbf{c}_j) \mathbf{c}_i^\top]$. Using the moment identity $h_\sigma(0) = \frac{\mathbb{E}[a_i]}{\sigma_i^2}$ and independence of the distribution μ over the $\mathbf{c}_j$, we can combine this term with the previous expectation, summing over boundary components, to obtain

$$\sum \frac{1}{\sigma_i^2} \mathbb{E} \left[a_i \left(f\left(\sum_j a_j \mathbf{c}_j \right) - f\left(\sum_{j\neq i} a_j \mathbf{c}_j \right) \right) \right] \mathbf{c}_i^\top. \tag{12}$$

A.3 Theorem 3

Theorem. *Let $\mathcal{M} \subset \mathbb{R}^d$ be a properly embedded manifold of positive codimension, μ a probability measure on $\mathcal{M}$, and $f : \mathbb{R}^d \to \mathbb{R}^m$ smooth with integrable derivative. Let $Q = \int_{\mathcal{M}} P_{N_x\mathcal{M}}\, d\mu$, and $K = \ker Q$. Then for any linear map $B : \mathbb{R}^d \to \mathbb{R}^m$ such that $BK = 0$, there is a smooth map $g : \mathbb{R}^d \to \mathbb{R}^m$ such that $g|_{\mathcal{M}} = f|_{\mathcal{M}}$, and $\int_{\mathcal{M}} Dg P_N\, d\mu = B + \int_{\mathcal{M}} Df P_N\, d\mu$. In particular, the Jacobian lens can be freely modified by any matrix B that annihilates $K = \{v \in \mathbb{R}^d \mid v \in T_x\mathcal{M}$ almost everywhere$\}$.*

We first restate the result from (Anders et al., 2020):

Theorem ((Anders et al., 2020)). *Let $\mathcal{M} \subset \mathbb{R}^d$ be a properly embedded manifold with positive codimension, and let $\{V_i\}_{1 \le i \le m}$ be conservative vector fields in $N\mathcal{M}$. Then there is a map $r : \mathbb{R}^d \to \mathbb{R}^m$ such that $r|_{\mathcal{M}} = 0$ and the gradient of r on $N\mathcal{M}$ is given by the V_i on $N\mathcal{M}$.*

Anders et al. (2020) constructs this for a scalar function and single V, using a partition of unity argument to perform the extensions locally; we refer to their paper for details.

To recover the statement of Theorem 3, let $Q^\dagger$ be the Moore-Penrose pseudoinverse of Q, and set the extensions V_i in Lemma A.3 to be $V_i(x) = P_{N_x\mathcal{M}} w_i$, where $w_i = (BQ^\dagger)^\top e_i$. Then the r constructed by the lemma satisfies

$$\int_{\mathcal{M}} D(f+r)P_N \, d\mu = \int_{\mathcal{M}} Df P_N \, d\mu + \int_{\mathcal{M}} BQ^\dagger P_N \, d\mu = \int_{\mathcal{M}} Df P_N \, d\mu + BQ^\dagger Q. \tag{13}$$

Since $Q^\dagger Q$ is projection onto $K^\perp$, and B annihilates K, the result follows.

To see that K contains only vectors that are in the tangent space almost everywhere, note that $Qv = 0$ implies $\langle v, Qv \rangle = 0$. Inserting the definition of Q, we have $\langle v, \int P_{N_x} v \, d\mu \rangle = \int \langle v, P_{N_x} v \rangle \, d\mu$. Since P_{N_x} is idempotent and self adjoint, we have this equals $\int \langle P_{N_x} v, P_{N_x} v \rangle \, d\mu$. Note that the argument is the norm of $P_{N_x} v$, which is strictly nonnegative, and 0 precisely when $v \in T_x\mathcal{M}$. So for the integral to be 0, v must be in $T_x\mathcal{M}$ μ-almost everywhere.

B Experiment details

B.1 Evaluation protocol

Throughout this paper, we evaluated semantic information content of vectors using the synthetic lens evaluation benchmarks of (Gurnee et al., 2026): at designated token positions, over a band of network layers, this measures the rank of task-specific tokens measured by the J-lens. Throughout, we report median minimum rank, in which we find the minimum rank over the band of measured layers and target tokens for each benchmark item, and report the median. Pass@k values are additionally reported in Appendix C. Gurnee et al. (2026)'s suite consists of the following tasks:

- **multihop**: In two-hop factual inferences, this measures whether the model has calculated the correct answer earlier in the sequence.
- **order-ops**: Scans for intermediate results in arithmetic calculations.
- **multilingual**: Calculates whether the model represents target language, and source and target concepts, in multi-lingual sequences.
- **poetry**: Checks for the existence of an appropriate rhyming word in poetic couplets.
- **typo**: Checks for the existence of a corrected token in sequences that contain a typo.
- **association**: Evaluates the presence of unstated abstract concepts, like emotions and genres, in the model's workspace.

Reported median ranks in the main document are the mean of the medians over the six tasks in the suite; we include breakdowns by evaluation in Section C.

B.2 Pushforward hyperparameters

For pushforward tangent spaces on all three models we used PCA on $k = 128$ nearest neighbors (computed by Euclidean distance) in H_0 to approximate a tangent space on $\mathbf{h}_{0,t-j}$ for $j \in \{0, 1, ..., 15\}$. We preserved 99% of the energy of the PCA decompositions, and 99% of the pushforward energy to form the pushforward tangent space—we did not restrict to a specific maximum or minimum dimension. The output horizon was the remainder of the sequence, $\mathbf{h}_{L,t}, \mathbf{h}_{L,t+1}, ..., \mathbf{h}_{L,t_{\text{final}}}$.

Qwen was evaluated at layers 8, 12, 16, 20, and 24 with target layer 27; OLMo used layers 8, 12, 16, 20, 24, and 28 with target layer 31; LLaMA used layers 8, 12, 16, 20, and 24 with target layer 27.

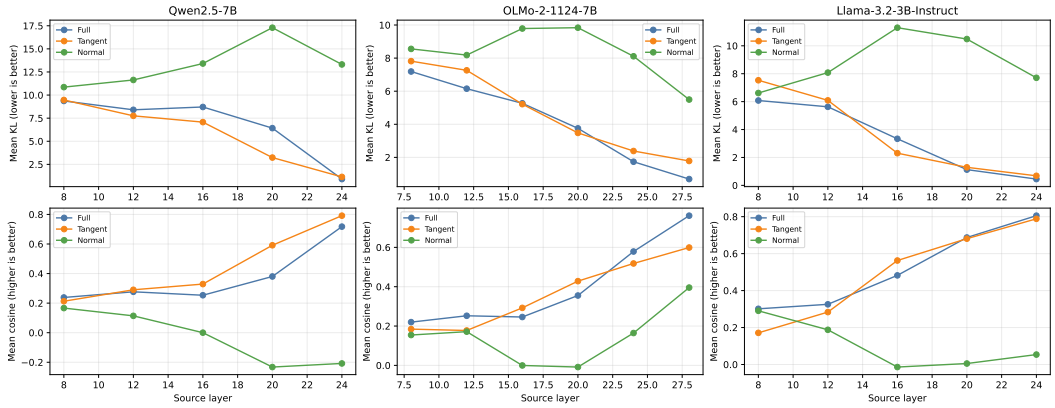

Figure 4: Layerwise average KL divergence (top) and cosine similarity (bottom) between the network output $f(\mathbf{h}_\ell)$ and the three linear approximations we study: $\mathbb{E}[J]\mathbf{h}_\ell$, $\mathbb{E}[J]P_T\mathbf{h}_\ell$, and $\mathbb{E}[J]P_N\mathbf{h}_\ell$.

To determine hyperparameters, we studied a range of different values on the first 30 indices of the evaluation suite described above. We then reported results exclusively on held out indices to prevent contamination of our results.

The global J-lens was calculated using 1000 sequences from the WikiText-103 dataset (Merity et al., 2016). For tangent projected lenses (Section 5.1), we decreased to 400 sequences, and calculated the lens contribution at only four tokens in each sequence to defray computation costs. We used identical sequences and positions to compute the compared full Jacobian and normal-projected Jacobian lenses.

B.3 PCA HYPERPARAMETERS

As with the pushforward, we used a 99% spectral energy threshold, using $k = 128$ nearest neighbors. We retrieved neighbors from a small corpus of 1000 sequences from the WikiText-103 dataset (Merity et al., 2016).

C EXTENDED RESULTS

C.1 LAYERWISE DECOMPOSITION OF THE TANGENT PROJECTED LATENTS

In addition to recording benchmark data, we tracked the KL divergence between $\mathbb{E}[J]h_\ell$ and the actual output of the transformer, h_L, when projected into token space. This gives a particularly clean tool to examine the decomposition of hidden states into tangent and normal components independently of the studied benchmarks. We report the layerwise breakdown for the pushforward tangent spaces in Figure 4. The tangent projection is most useful, for all three models, specifically in the mid-to-late layers most associated with the linear representation hypothesis. In Qwen, the denoising effect observed in Section 4.1 is visible in the way that this advantage is preserved in late layers of the network.

C.2 EXACT LOCAL JACOBIANS

To study the effect averaging has $\int_M \frac{\partial h_L}{\partial h_\ell} d\mu$ on the Jacobian lens, we additionally recorded results when, instead, the exact local Jacobian is used at each point on the activation manifold, shown in Table 4. For ease of reference, we reproduce the results of Tables 1.

Table 4: Local Jacobian results for Qwen, OLMo, and LLaMA, with expected Jacobian results for comparison.

Model	Variant	P@1 ↑	P@5 ↑	P@10 ↑	Mean med. rank ↓	KL ↓	Cos. ↑
Qwen-2.5 7B	$\mathbb{E}[J]\mathbf{h}_\ell$	0.195	0.356	0.407	461.8	10.429	0.405
	$\mathbb{E}[J]P_T(\mathbf{h}_\ell)$	0.267	0.437	0.546	137.0	9.286	0.470
	$\mathbb{E}[J]P_N(\mathbf{h}_\ell)$	0.002	0.011	0.011	9609.4	15.801	-0.098
	$J_{\mathbf{h}_\ell}\mathbf{h}_\ell$	0.295	0.397	0.414	466.2	6.762	0.373
	$J_{\mathbf{h}_\ell}P_T(\mathbf{h}_\ell)$	0.236	0.367	0.391	616.2	5.735	0.443
	$J_{\mathbf{h}_\ell}P_N(\mathbf{h}_\ell)$	0.012	0.048	0.085	2970.4	13.302	-0.032
OLMo-2 7B	$\mathbb{E}[J]\mathbf{h}_\ell$	0.204	0.395	0.507	251.3	7.299	0.485
	$\mathbb{E}[J]P_T(\mathbf{h}_\ell)$	0.187	0.363	0.419	169.5	7.801	0.422
	$\mathbb{E}[J]P_N(\mathbf{h}_\ell)$	0.142	0.277	0.352	528.1	7.899	0.366
	$J_{\mathbf{h}_\ell}\mathbf{h}_\ell$	0.286	0.359	0.393	460.9	4.132	0.402
	$J_{\mathbf{h}_\ell}P_T(\mathbf{h}_\ell)$	0.290	0.318	0.349	590.1	4.657	0.367
	$J_{\mathbf{h}_\ell}P_N(\mathbf{h}_\ell)$	0.041	0.086	0.101	871.0	8.328	0.146
LLaMA-3.2 3B	$\mathbb{E}[J]\mathbf{h}_\ell$	0.305	0.480	0.604	23.9	4.719	0.606
	$\mathbb{E}[J]P_T(\mathbf{h}_\ell)$	0.320	0.473	0.571	19.4	4.842	0.573
	$\mathbb{E}[J]P_N(\mathbf{h}_\ell)$	0.065	0.134	0.208	278.5	7.453	0.263
	$J_{\mathbf{h}_\ell}\mathbf{h}_\ell$	0.238	0.430	0.483	187.3	3.325	0.520
	$J_{\mathbf{h}_\ell}P_T(\mathbf{h}_\ell)$	0.237	0.436	0.483	369.0	3.581	0.497
	$J_{\mathbf{h}_\ell}P_N(\mathbf{h}_\ell)$	0.082	0.166	0.199	572.3	8.837	0.105

Table 5: Median best ranks for the three models, divided by evaluation suite, on held-out data. Bolded are the best global mean and best local Jacobian approximation for each suite in each model family.

Model	Projection	Multihop	Order-Ops	Multilingual	Typo	Poetry	Association
Qwen-2.5 7B	$\mathbb{E}[J]\mathbf{h}_\ell$	5	35	6	65.5	**1.5**	2658
	$\mathbb{E}[J]P_T(\mathbf{h}_\ell)$	**3**	**6**	**5**	**30.5**	**1.5**	**776**
	$\mathbb{E}[J]P_N(\mathbf{h}_\ell)$	12723	12942	704	3923.5	6471	20893
	$J_{\mathbf{h}_\ell}\mathbf{h}_\ell$	45	**1**	**53**	194.5	**1**	**2502.5**
	$J_{\mathbf{h}_\ell}P_T(\mathbf{h}_\ell)$	**25**	**1**	64	**130**	**1**	3476.5
	$J_{\mathbf{h}_\ell}P_N(\mathbf{h}_\ell)$	4538	566	395	3417.5	673	8233
OLMo-2 7B	$\mathbb{E}[J]\mathbf{h}_\ell$	**15**	7	9	14.5	**1**	1461.5
	$\mathbb{E}[J]P_T(\mathbf{h}_\ell)$	29	26	**7**	**9**	**1**	**945**
	$\mathbb{E}[J]P_N(\mathbf{h}_\ell)$	19	**3**	48	165.5	2	2931
	$J_{\mathbf{h}_\ell}\mathbf{h}_\ell$	**74**	**1**	**89**	**353**	**1**	**2247.5**
	$J_{\mathbf{h}_\ell}P_T(\mathbf{h}_\ell)$	179	**1**	247	725	**1**	2387.5
	$J_{\mathbf{h}_\ell}P_N(\mathbf{h}_\ell)$	162	74	192	578	82	4138
LLaMA-3.2 3B	$\mathbb{E}[J]\mathbf{h}_\ell$	**16**	**7**	**3**	2	**1**	114.5
	$\mathbb{E}[J]P_T(\mathbf{h}_\ell)$	23	17	**3**	**1**	**1**	**71.5**
	$\mathbb{E}[J]P_N(\mathbf{h}_\ell)$	104	16	41	539.5	47	923.5
	$J_{\mathbf{h}_\ell}\mathbf{h}_\ell$	**18**	3	**23**	98.5	**1**	**980.5**
	$J_{\mathbf{h}_\ell}P_T(\mathbf{h}_\ell)$	**18**	**2**	46	**70.5**	**1**	2076.5
	$J_{\mathbf{h}_\ell}P_N(\mathbf{h}_\ell)$	553	395	28	959	6	1493

C.3 Local projection results by evaluation suite

We decompose the macroscopic results into their suite-by-suite decomposition in Table 5. Qwen shows that the tangent projection is a clean improvement on the full Jacobian across all suites, and LLaMA similarly shows very comparable performance between the tangent and full Jacobians. OLMo shows more mixed results. One confounding factor, which likely damages the performance of the exact local Jacobians $J_{\mathbf{h}_\ell}$, is that the Jacobian isn't determined purely by the network weights, but also by the specific contextual information, which adds noise to the signal.

Table 6: Held-out evaluation results for Qwen, OLMo, and LLaMA with direct PCA tangent spaces, including both using the global approximator $\mathbb{E}[J]$ and exact local Jacobians $J_{\mathbf{h}_\ell}$.

Model	Variant	P@1	P@5	P@10	Mean med. rank	KL ↓	Cos. ↑
Qwen-2.5 7B	$\mathbb{E}[J]\mathbf{h}_\ell$	0.195	0.356	0.407	461.8	10.429	0.405
	$\mathbb{E}[J]P_T(\mathbf{h}_\ell)$	0.205	0.379	0.439	193	10.264	0.437
	$\mathbb{E}[J]P_N(\mathbf{h}_\ell)$	0.000	0.000	0.000	14437.6	17.148	-0.053
	$J_{\mathbf{h}_\ell}\mathbf{h}_\ell$	0.295	0.397	0.414	466.2	6.762	0.373
	$J_{\mathbf{h}_\ell}P_T(\mathbf{h}_\ell)$	0.229	0.326	0.379	1052.2	6.541	0.391
	$J_{\mathbf{h}_\ell}P_N(\mathbf{h}_\ell)$	0.022	0.049	0.056	3095.7	13.801	-0.040
OLMo-2 7B	$\mathbb{E}[J]\mathbf{h}_\ell$	0.204	0.395	0.507	251.3	7.299	0.485
	$\mathbb{E}[J]P_T(\mathbf{h}_\ell)$	0.199	0.372	0.496	231.6	7.385	0.483
	$\mathbb{E}[J]P_N(\mathbf{h}_\ell)$	0.000	0.000	0.000	3532.3	9.374	0.173
	$J_{\mathbf{h}_\ell}\mathbf{h}_\ell$	0.286	0.359	0.393	460.9	4.132	0.402
	$J_{\mathbf{h}_\ell}P_T(\mathbf{h}_\ell)$	0.281	0.350	0.384	440.2	4.014	0.423
	$J_{\mathbf{h}_\ell}P_N(\mathbf{h}_\ell)$	0.000	0.033	0.058	2739.5	11.973	-0.133
LLaMA-3.2 3B	$\mathbb{E}[J]\mathbf{h}_\ell$	0.305	0.480	0.604	23.9	4.719	0.606
	$\mathbb{E}[J]P_T(\mathbf{h}_\ell)$	0.301	0.447	0.519	28.7	5.299	0.576
	$\mathbb{E}[J]P_N(\mathbf{h}_\ell)$	0.145	0.222	0.233	576.3	7.514	0.290
	$J_{\mathbf{h}_\ell}\mathbf{h}_\ell$	0.238	0.430	0.483	187.3	3.325	0.520
	$J_{\mathbf{h}_\ell}P_T(\mathbf{h}_\ell)$	0.216	0.346	0.393	437.8	4.048	0.419
	$J_{\mathbf{h}_\ell}P_N(\mathbf{h}_\ell)$	0.101	0.214	0.279	271.2	9.666	0.086

C.4 Exact local Jacobians with PCA tangent spaces

We likewise recorded the local projection results with PCA approximated tangent spaces, and report them in Table 6. As above, we duplicate the projection of the global results here for easier comparison.

C.5 Tangent- and normal-projected Jacobians by evaluation suite

We give a breakdown of the results reported in Table 3 for each model family, divided by evaluation suite, in Tables 7–9. Unlike in the main body, here for overall median rank we report pooled median, rather than the unweighted average over suites.

C.6 Layerwise coherence and Frobenius statistics for projected lenses

In Figure 3, we summarized the cancellation of the tangent projected Jacobians in terms of relative coherence. In Tables 10–12, we provide extended results from this study, including tracking the local Frobenius norm change, which tracks precisely with the relative rank of the approximated tangent space, and the Frobenius norm of the global objects, which substantially decreases (local projections lose coherence) as layer index increases.

Table 7: Breakdown by evaluation suite for Qwen-2.5 7B of the J-lens tangent and normal projections

Benchmark	Lens	Pass@1	Pass@5	Pass@10	Mean KL	Median rank
Multihop ($n = 63$)	Logit lens	12.7	22.2	25.4	9.10	196
	$\mathbb{E}[J]$	30.2	54.0	57.1	8.94	5
	$\mathbb{E}[JP_T]$	22.2	42.9	54.0	9.82	10
	$\mathbb{E}[JP_N]$	6.3	15.9	28.6	10.10	34
Order-ops ($n = 25$)	Logit lens	28.0	56.0	68.0	7.18	3
	$\mathbb{E}[J]$	16.0	24.0	32.0	12.96	35
	$\mathbb{E}[JP_T]$	20.0	40.0	52.0	13.46	7
	$\mathbb{E}[JP_N]$	0.0	0.0	4.0	13.89	269
Multilingual ($n = 77$)	Logit lens	27.3	48.1	51.9	10.30	8
	$\mathbb{E}[J]$	23.4	51.9	54.5	10.35	5
	$\mathbb{E}[JP_T]$	5.2	6.5	9.1	11.51	93
	$\mathbb{E}[JP_N]$	26.0	50.6	55.8	10.65	5
Typo ($n = 66$)	Logit lens	3.0	15.2	19.7	10.73	77
	$\mathbb{E}[J]$	1.5	3.0	10.6	10.44	80
	$\mathbb{E}[JP_T]$	0.0	1.5	4.5	9.48	223.5
	$\mathbb{E}[JP_N]$	1.5	10.6	16.7	12.14	81
Poetry ($n = 68$)	Logit lens	42.6	61.8	73.5	10.05	2
	$\mathbb{E}[J]$	50.0	80.9	83.8	10.64	1.5
	$\mathbb{E}[JP_T]$	35.3	57.4	64.7	12.83	3.5
	$\mathbb{E}[JP_N]$	41.2	72.1	79.4	10.85	2
Association ($n = 72$)	Logit lens	0.0	0.0	0.0	7.97	2,988.5
	$\mathbb{E}[J]$	0.0	0.0	0.0	8.51	2,233
	$\mathbb{E}[JP_T]$	0.0	0.0	0.0	10.84	1,520
	$\mathbb{E}[JP_N]$	1.4	1.4	2.8	8.52	3,658
Overall ($n = 371$)	Logit lens	18.1	31.5	36.7	9.46	39
	$\mathbb{E}[J]$	20.5	36.9	40.4	10.00	24
	$\mathbb{E}[JP_T]$	12.7	22.1	27.2	11.10	84
	$\mathbb{E}[JP_N]$	14.6	28.6	34.8	10.66	46

Table 8: Breakdown by evaluation suite for OLMo-2 7B of the J-lens, as well as its tangent and normal projections.

Benchmark	Lens	Pass@1	Pass@5	Pass@10	Mean KL	Median rank
Multihop ($n = 63$)	Logit lens	12.7	30.2	34.9	6.03	38
	$\mathbb{E}[J]$	14.3	28.6	36.5	7.91	28
	$\mathbb{E}[JP_T]$	6.3	15.9	22.2	7.55	51
	$\mathbb{E}[JP_N]$	15.9	27.0	36.5	8.60	25
Order-ops ($n = 25$)	Logit lens	68.0	80.0	80.0	5.82	1
	$\mathbb{E}[J]$	0.0	12.0	44.0	9.32	11
	$\mathbb{E}[JP_T]$	0.0	0.0	4.0	8.99	38
	$\mathbb{E}[JP_N]$	0.0	12.0	28.0	9.81	18
Multilingual ($n = 77$)	Logit lens	27.3	58.4	70.1	6.86	3
	$\mathbb{E}[J]$	23.4	37.7	44.2	6.42	20
	$\mathbb{E}[JP_T]$	3.9	13.0	20.8	6.67	78
	$\mathbb{E}[JP_N]$	24.7	33.8	44.2	6.96	20
Typo ($n = 66$)	Logit lens	4.5	19.7	30.3	6.25	31
	$\mathbb{E}[J]$	22.7	57.6	72.7	7.79	3
	$\mathbb{E}[JP_T]$	0.0	1.5	3.0	6.61	964
	$\mathbb{E}[JP_N]$	28.8	69.7	77.3	8.89	2
Poetry ($n = 68$)	Logit lens	79.4	92.6	94.1	5.61	1
	$\mathbb{E}[J]$	60.3	83.8	88.2	9.83	1
	$\mathbb{E}[JP_T]$	23.5	48.5	54.4	10.39	6.5
	$\mathbb{E}[JP_N]$	58.8	83.8	88.2	10.10	1
Association ($n = 72$)	Logit lens	1.4	6.9	9.7	5.56	526.5
	$\mathbb{E}[J]$	0.0	0.0	1.4	7.54	1,537.5
	$\mathbb{E}[JP_T]$	0.0	0.0	0.0	8.43	4,552
	$\mathbb{E}[JP_N]$	0.0	1.4	1.4	7.54	1,023.5
Overall ($n = 371$)	Logit lens	28.0	44.5	50.4	6.06	10
	$\mathbb{E}[J]$	22.4	39.1	47.7	7.96	14
	$\mathbb{E}[JP_T]$	6.2	14.6	18.9	7.99	125
	$\mathbb{E}[JP_N]$	23.7	40.4	47.4	8.46	13

Table 9: Breakdown by evaluation suite for LLaMA-3.2 3B of the Jacobian, as well as its intrinsic tangent and extrinsic normal projections.

Benchmark	Lens	Pass@1	Pass@5	Pass@10	Mean KL	Median rank
Multihop ($n = 63$)	Logit lens	7.9	17.5	23.8	5.82	69
	$\mathbb{E}[J]$	14.3	33.3	36.5	4.32	23
	$\mathbb{E}[JP_T]$	6.3	19.0	28.6	4.48	32
	$\mathbb{E}[JP_N]$	14.3	31.7	38.1	5.36	21
Order-ops ($n = 25$)	Logit lens	8.0	36.0	36.0	6.73	18
	$\mathbb{E}[J]$	0.0	32.0	64.0	4.36	8
	$\mathbb{E}[JP_T]$	8.0	36.0	60.0	4.33	8
	$\mathbb{E}[JP_N]$	0.0	12.0	28.0	5.01	17
Multilingual ($n = 77$)	Logit lens	40.3	53.2	63.6	7.10	3
	$\mathbb{E}[J]$	42.9	66.2	75.3	6.50	2
	$\mathbb{E}[JP_T]$	11.7	37.7	46.8	6.77	13
	$\mathbb{E}[JP_N]$	44.2	63.6	71.4	6.93	2
Typo ($n = 66$)	Logit lens	60.6	81.8	87.9	6.13	1
	$\mathbb{E}[J]$	47.0	65.2	69.7	4.56	2
	$\mathbb{E}[JP_T]$	0.0	0.0	4.5	4.40	110
	$\mathbb{E}[JP_N]$	57.6	68.2	71.2	6.15	1
Poetry ($n = 68$)	Logit lens	80.9	97.1	97.1	5.63	1
	$\mathbb{E}[J]$	72.1	89.7	97.1	4.66	1
	$\mathbb{E}[JP_T]$	55.9	82.4	94.1	6.00	1
	$\mathbb{E}[JP_N]$	67.6	89.7	97.1	4.91	1
Association ($n = 72$)	Logit lens	1.4	5.6	9.7	6.11	684.5
	$\mathbb{E}[J]$	4.2	6.9	16.7	3.70	87.5
	$\mathbb{E}[JP_T]$	0.0	0.0	0.0	4.57	389.5
	$\mathbb{E}[JP_N]$	4.2	9.7	15.3	4.80	155
Overall ($n = 371$)	Logit lens	36.1	49.9	55.0	6.23	6
	$\mathbb{E}[J]$	33.7	50.9	59.6	4.76	5
	$\mathbb{E}[JP_T]$	14.3	28.6	36.7	5.23	34
	$\mathbb{E}[JP_N]$	35.0	49.9	56.6	5.61	6

Table 10: Rank, directional-energy, and coherence decomposition for Qwen-2.5 7B. Relative rank is the norm ratio $\sqrt{\text{rank}/d}$ weighted by per-sample full-Jacobian energy; relative mean Frobenius norm is the ratio of local Frobenius norms $\sqrt{\mathbb{E}\|JP_T\|_F^2/\mathbb{E}\|J\|_F^2}$; Relative Frobenius norm is $\|\mathbb{E}[JP_T]\|_F/\|\mathbb{E}[J]\|_F$; and relative coherence is the ratio of the tangent to total coherence, where coherence is $\frac{||\frac{1}{n}\sum J_x P_x||_F^2}{\frac{1}{n}\sum ||J_x P_x||_F^2}$, where $P_x = P_{T_x}$ for the tangent projection, and I for the full map.

Layer	Relative rank	Relative Mean Frob. Norm	Relative Frob. Norm	Relative Coherence
4	0.407	0.421	0.336	0.637
8	0.400	0.448	0.378	0.715
12	0.375	0.401	0.376	0.880
16	0.363	0.427	0.355	0.690
20	0.337	0.511	0.296	0.334
24	0.304	0.464	0.164	0.126
26	0.298	0.322	0.119	0.136

Table 11: Rank, directional-energy, and coherence decomposition for OLMo-2 7B. Relative rank is the norm ratio $\sqrt{\text{rank}/d}$ weighted by per-sample full-Jacobian energy; relative mean Frobenius norm is the ratio of local Frobenius norms $\sqrt{\mathbb{E}\|JP_T\|_F^2/\mathbb{E}\|J\|_F^2}$; Relative Frobenius norm is $\|\mathbb{E}[JP_T]\|_F/\|\mathbb{E}[J]\|_F$; and relative coherence is the ratio of the tangent to total coherence, where coherence is $\frac{||\frac{1}{n}\sum J_x P_x||_F^2}{\frac{1}{n}\sum ||J_x P_x||_F^2}$, where $P_x = P_{T_x}$ for the tangent projection, and I for the full map.

Layer	Relative rank	Relative Mean Frob. Norm	Relative Frob. Norm	Relative Coherence
4	0.319	0.352	0.271	0.590
8	0.340	0.399	0.335	0.706
12	0.294	0.368	0.305	0.689
16	0.233	0.294	0.257	0.767
20	0.200	0.228	0.197	0.747
24	0.199	0.215	0.167	0.604
28	0.182	0.191	0.127	0.443
30	0.161	0.154	0.104	0.461

Table 12: Rank, directional-energy, and coherence decomposition for LLaMA-3.2 3B. Relative rank is the norm ratio $\sqrt{\text{rank}/d}$ weighted by per-sample full-Jacobian energy; relative mean Frobenius norm is the ratio of local Frobenius norms $\sqrt{\mathbb{E}\|JP_T\|_F^2/\mathbb{E}\|J\|_F^2}$; Relative Frobenius norm is $\|\mathbb{E}[JP_T]\|_F/\|\mathbb{E}[J]\|_F$; and relative coherence is the ratio of the tangent to total coherence, where coherence is $\frac{||\frac{1}{n}\sum J_x P_x||_F^2}{\frac{1}{n}\sum ||J_x P_x||_F^2}$, where $P_x = P_{T_x}$ for the tangent projection, and I for the full map.

Layer	Relative rank	Relative Mean Frob. Norm	Relative Frob. Norm	Relative Coherence
4	0.358	0.402	0.326	0.660
8	0.390	0.481	0.370	0.593
12	0.369	0.497	0.351	0.498
16	0.314	0.395	0.210	0.282
20	0.281	0.330	0.132	0.160
24	0.277	0.297	0.105	0.126
26	0.285	0.284	0.105	0.136