

Predictive Geometry of Hidden Trajectories in Transformers

Timur Mudarisov¹ Mikhail Burtsev² Tatiana Petrova¹ Radu State¹

¹University of Luxembourg

²London Institute for Mathematical Sciences

NeurIPS 2026 · Poster session, Paris · 11 December 2026



Paper repository

Outline

Motivation and contributions

Problem setting

Theory

Empirical validation

Follow-up analyses

Discussion

Section 1

Motivation and contributions

Motivation and contributions

Problem setting

Theory

Empirical validation

Follow-up analyses

Discussion

Motivation

Decoder-only transformers [21, 4] are trained *only* through a terminal next-token loss, yet this loss constrains every intermediate hidden state through the fixed downstream computation.

Question

Near hidden-state trajectories on which a trained model already predicts well, which residual-stream directions are constrained by the terminal loss, and which are (approximately) invisible to it?

Prior evidence of structure

- intrinsic dimension of hidden representations follows an expansion–contraction profile across layers [20];
- activation-space low-rank compression preserves performance [3, 23, 18];
- token pruning is typically guided by attention [25, 14, 24].

Gap. This structure is characterised empirically; it is not derived from the prediction objective and gives no loss-aware notion of which directions or tokens matter.

Contributions

1. **Loss-induced hidden geometry.** For a fixed trained decoder we define layerwise loss-to-go functions $\mathcal{J}_\ell$ and show that, near successful trajectories, their curvature is a pullback Fisher operator $K_\ell = \bar{J}_\ell^\top F^* \bar{J}_\ell$ up to a remainder of order $\sqrt{\varepsilon_*}$ (Theorem 1).
2. **Observable and prediction-null directions.** The spectrum of K_ℓ yields a testable local decomposition of hidden-state perturbations.
3. **Tokenwise predictive saliency.** A Fisher–Jacobian score $\kappa_{\ell,s}$ that vanishes outside the causal ancestors of the target and is bounded by downstream Jacobian couplings (Proposition 1).
4. **Geometry-guided reduction and recovery.** Ten LLMs (1B–9B), three corpora: perturbation sensitivity, rank allocation, token pruning, low-rank distillation.

Section 2

Problem setting

Motivation and contributions

Problem setting

Theory

Empirical validation

Follow-up analyses

Discussion

Setting and notation

Fixed trained decoder-only pre-norm transformer with L blocks, hidden size d , vocabulary $\mathcal{V}$; fixed target position t .

$$X_\ell \in \mathcal{X}_t := \mathbb{R}^{t \times d}, \quad X_{\ell+1} = B_\ell(X_\ell), \quad \Phi_{\ell \rightarrow L} := B_{L-1} \circ \cdots \circ B_\ell,$$

$$\Psi_\ell(X) := Z_t(\Phi_{\ell \rightarrow L}(X)) \in \mathbb{R}^{|\mathcal{V}|} \text{ (target logits)}, \quad \phi(z) := D_{\text{KL}}(q \parallel \text{softmax}(z)).$$

Definition 1 (Loss-to-go and low-loss sets)

$$\mathcal{J}_\ell(X) := \phi(\Psi_\ell(X)), \quad \mathcal{A}_\varepsilon^\ell := \{X \in \mathcal{X}_t : \mathcal{J}_\ell(X) \leq \varepsilon\} = \Phi_{\ell \rightarrow L}^{-1}(\mathcal{A}_\varepsilon^L).$$

$\mathcal{A}_\varepsilon^\ell$ collects the layer- ℓ states that the fixed downstream network still maps to terminal loss $\leq \varepsilon$. We study the *local geometry* of $\mathcal{A}_\varepsilon^\ell$ around a successful reference trajectory $X_0^* \rightarrow \cdots \rightarrow X_L^*$ (forward dynamics of a trained model, not training dynamics).

Softmax Fisher geometry

Lemma 1 (Softmax Fisher identities)

For $p = \text{softmax}(z)$,

$$\nabla_z \phi(z) = p - q, \quad \nabla_z^2 \phi(z) = F(p) := \text{diag}(p) - pp^\top, \quad F(p)\mathbf{1} = 0.$$

- The Hessian of the KL loss on logits is the Fisher information of the categorical output distribution.
- $F(p)\mathbf{1} = 0$: the logit-shift direction is invisible to the softmax. We therefore work with centred logits

$$P := I - \frac{1}{|\mathcal{V}|} \mathbf{1}\mathbf{1}^\top, \quad \bar{\Psi}_\ell := P\Psi_\ell, \quad \bar{J}_\ell := D\bar{\Psi}_\ell(X_\ell^*) \in \mathbb{R}^{|\mathcal{V}| \times td}.$$

- Reference quantities: $p^* := \text{softmax}(\Psi_\ell(X_\ell^*))$, $F^* := F(p^*)$, $\varepsilon_* := \mathcal{J}_\ell(X_\ell^*)$.

Section 3

Theory

Motivation and contributions

Problem setting

Theory

Empirical validation

Follow-up analyses

Discussion

Main result: local output geometry of low-loss sets

Theorem 1

Let $\Psi_\ell \in C^2$ near X_ℓ^* with $(\sum_a \|\nabla^2 \Psi_{\ell,a}(X)\|_{\text{op}}^2)^{1/2} \leq M_\ell$. Then:

1. *Hessian decomposition.* $\nabla^2 \mathcal{J}_\ell(X_\ell^*) = \bar{J}_\ell^\top F^* \bar{J}_\ell + R_\ell$,
 $\|R_\ell\|_{\text{op}} \leq M_\ell \|p^* - q\|_2 \leq M_\ell \sqrt{2\varepsilon_*}$.
2. *Pullback Fisher geometry.* With $K_\ell := \bar{J}_\ell^\top F^* \bar{J}_\ell$ and $g_\ell(\delta X, \delta X) := \langle \delta X, K_\ell \delta X \rangle$,
$$\mathcal{J}_\ell(X_\ell^* + \delta X) = \mathcal{J}_\ell(X_\ell^*) + \langle \nabla \mathcal{J}_\ell(X_\ell^*), \delta X \rangle + \frac{1}{2} g_\ell(\delta X, \delta X) + O(\sqrt{\varepsilon_*} \|\delta X\|^2 + \|\delta X\|^3).$$
3. *Prediction-null directions.* If $\bar{J}_\ell$ has constant rank near X_ℓ^* , then $\mathcal{N}_\ell := \ker \bar{J}_\ell$ consists of first-order prediction-null directions.

The curvature bound holds locally for smooth activations and LayerNorm where LayerNorm denominators stay away from zero and attention probabilities remain finite.

Proof sketch of Theorem 1

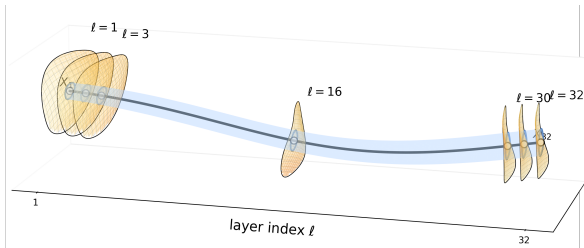
1. Softmax is shift-invariant, so $\mathcal{J}_\ell = \phi \circ \bar{\Psi}_\ell$. By the second-order chain rule and Lemma 1,

$$\nabla^2 \mathcal{J}_\ell(X_\ell^*) = \bar{J}_\ell^\top F^* \bar{J}_\ell + R_\ell, \quad R_\ell = \sum_a (p_a^* - q_a) \nabla^2 \bar{\Psi}_{\ell,a}(X_\ell^*).$$

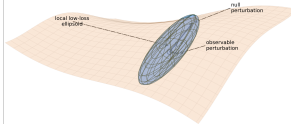
2. Cauchy–Schwarz gives $\|R_\ell\|_{\text{op}} \leq M_\ell \|p^* - q\|_2$, and Pinsker gives $\|p^* - q\|_2 \leq \|p^* - q\|_1 \leq \sqrt{2D_{\text{KL}}(q\|p^*)} = \sqrt{2\varepsilon_*}$.
3. Second-order Taylor expansion with bounded third derivative; the remainder contributes $|\langle \delta X, R_\ell \delta X \rangle| \leq M_\ell \sqrt{2\varepsilon_*} \|\delta X\|^2$.
4. For $v \in \ker \bar{J}_\ell$ the centred target logits are unchanged to first order; constant rank yields a local foliation by prediction-null leaves.

Interpretation. A successful trajectory lies in a tube of low-loss states: thin along output-sensitive directions (large eigenvalues of K_ℓ), thick along $\mathcal{N}_\ell$.

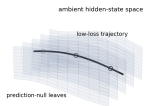
Geometric picture



Local pullback geometry around a low-loss point



3D prediction-null foliation and observable quotient



Top: hidden-state trajectory and layerwise low-loss sets $\mathcal{A}_\varepsilon^\ell$. **Bottom left:** local pullback geometry near X_ℓ^* . **Bottom right:** prediction-null foliation and observable quotient.

Tokenwise pullback saliency

Let $P_s : \mathcal{X}_t \rightarrow \mathbb{R}^d$ project onto token $s \leq t$, let $G_m^{(r \leftarrow u)}$ be the token blocks of $DB_m(X_m^*)$, and

$$H_{\ell,s} := D_{X_{\ell,s}} \bar{\Psi}_\ell(X_\ell^*) = \bar{J}_\ell P_s^\top \in \mathbb{R}^{|\mathcal{V}| \times d}, \quad \kappa_{\ell,s} := \text{tr}(P_s K_\ell P_s^\top).$$

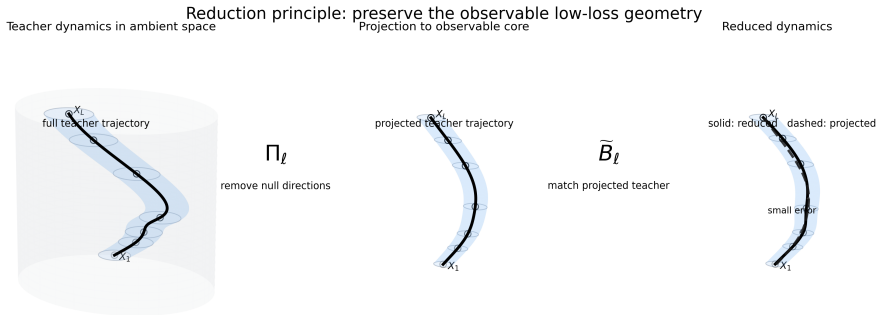
Proposition 1 (Tokenwise saliency and causal support)

1. *Fisher–Jacobian identity:* $\kappa_{\ell,s} = \text{tr}(H_{\ell,s}^\top F^* H_{\ell,s}) = \|F^{*1/2} H_{\ell,s}\|_F^2$.
2. *Causal support:* under the causal mask, $s \notin \text{Anc}_{\ell:L-1}(t) \Rightarrow H_{\ell,s} = 0, \kappa_{\ell,s} = 0$.
3. *Pathwise coupling bound:* with $\Pi_{\ell:s \rightarrow t}$ the causal paths from (ℓ, s) to (L, t) ,

$$\kappa_{\ell,s} \leq d \|F^*\|_{\text{op}} \|DZ_t(X_L^*)\|_{\text{op}}^2 \left(\sum_{\pi \in \Pi_{\ell:s \rightarrow t}} \prod_{(m,u \rightarrow r) \in \pi} \|G_m^{(r \leftarrow u)}\|_{\text{op}} \right)^2.$$

Geometric importance is *sparse* (exact zeros off the ancestor set) and *graded* (weak couplings imply small κ). Experiments use the measured $\kappa_{\ell,s}$, not the bound.

Geometry-guided reduction



Proposition 2 (Local loss guarantee)

If $\Pi_\ell^{(k)}$ projects onto the top- k eigenvectors of K_ℓ and $\delta X_{\text{tail}} = (I - \Pi_\ell^{(k)})\delta X$, then

$$\frac{1}{2} \langle \delta X_{\text{tail}}, K_\ell \delta X_{\text{tail}} \rangle \leq \frac{1}{2} \lambda_{k+1}(K_\ell) \|\delta X_{\text{tail}}\|^2.$$

Section 4

Empirical validation

Motivation and contributions

Problem setting

Theory

Empirical validation

Follow-up analyses

Discussion

Experimental protocol

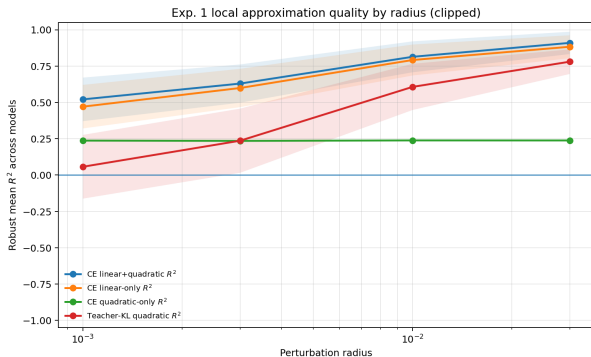
Model	Params
LLaMA-3.2-1B [7]	1B
SmolLM2-1.7B [2]	1.7B
Gemma-2B [5]	2B
Qwen2.5-3B [22]	3B
Phi-3 Mini [1]	3.8B
LLaMA-2-7B [19]	7B
Mistral-7B [12]	7B
Gemma-7B [5]	7B
OLMo-3-7B [16]	7B
Gemma-2-9B [5]	9B

Data. Held-out WikiText [15], OpenWebText [6], FineWeb [17]; disjoint calibration and evaluation sets; targets drawn from the second half of the context.

Experiments.

0. Local Taylor validation of $\frac{1}{2}\delta X^\top K_\ell \delta X$.
1. Observable/null anisotropy and $r_{\text{eff}}(K_\ell)$.
2. Spectral rank allocation vs. matched uniform SVD / LoRA [11, 9].
3. Token-update pruning by $\kappa_{\ell,s}$.
4. Geometry-regularised low-rank distillation.

Experiment 0: local Taylor validation



Model-averaged R^2 between true and predicted loss change vs. perturbation radius (clipped to $[-1, 1]$).

Perturb X_ℓ^* , propagate through the fixed downstream map, and compare with

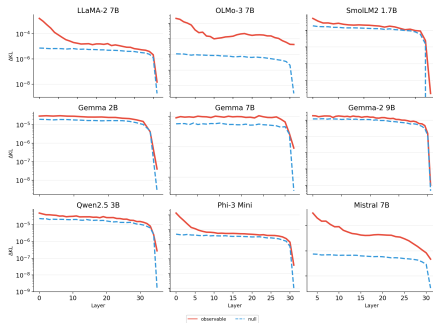
$$\Delta \text{CE}_{\text{lin}} = \langle \nabla_{X_\ell} \text{CE}, \delta X \rangle,$$

$$\Delta \text{CE}_{\text{quad}} = \frac{1}{2} \delta X^\top K_\ell \delta X.$$

- Supervised CE is *first-order dominated*: $\text{lin} \approx \text{lin} + \text{quad}$; quad-only $R^2 \approx 0.24$, since one-hot CE is not stationary at p^* .
- Teacher-KL ($q = p^*$) isolates the curvature: $\Delta D_{\text{KL}} \approx \frac{1}{2} \delta X^\top K_\ell \delta X$ reaches $R^2 \approx 0.6$ – 0.8 at moderate radii.

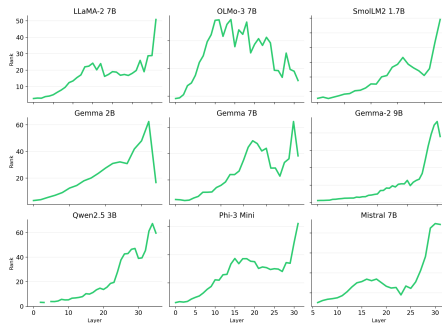
Experiment 1: anisotropy and effective observable dimension

Exp1: per-layer ΔK_L across models (wikitext)



ΔD_{KL} of equal-norm observable (red) vs. null (blue) perturbations.

Exp1: per-layer effective rank across models (wikitext)

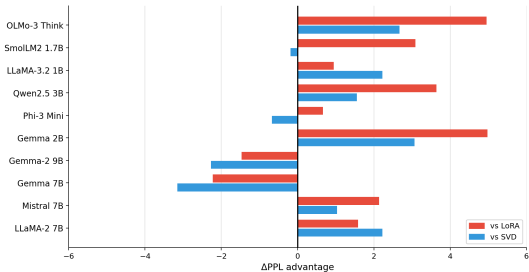


Effective observable rank $r_{\text{eff}}(K_\ell)$ across layers.

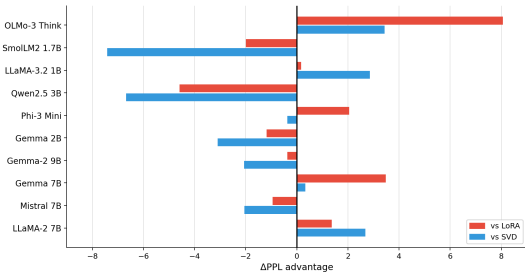
Observable perturbations are up to orders of magnitude more harmful than null ones at almost every layer (WikiText, nine models); $r_{\text{eff}}(K_\ell) = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ grows with depth but stays far below d .

Experiment 2: geometry-guided rank allocation

Geometry-Guided Δ PPL advantage at $\tau=0.05$ (wikitext)
Positive = geometry-guided wins | negative = baseline wins

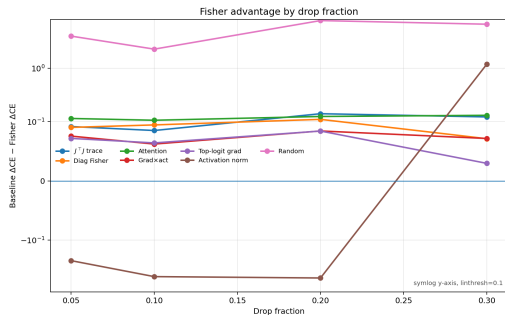
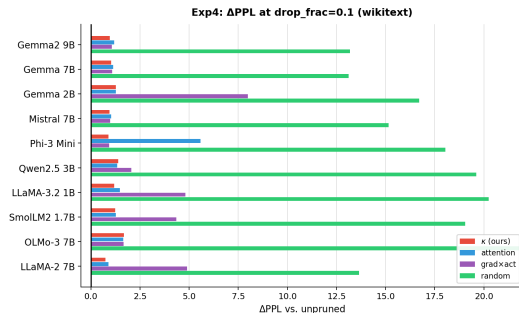


Geometry-Guided Δ PPL advantage at $\tau=0.1$ (wikitext)
Positive = geometry-guided wins | negative = baseline wins



Rank k_ℓ : smallest k with $\sum_{i>k} \lambda_i(K_\ell) \leq \tau \sum_i \lambda_i(K_\ell)$; $\Delta\text{PPL}_{\text{adv}} := \Delta\text{PPL}_{\text{uniform}} - \Delta\text{PPL}_{\text{geom}}$ at a matched budget (> 0 : geometry wins), WikiText, $\tau \in \{0.05, 0.1\}$. Gains are largest at moderate τ (Gemma-2B, Qwen2.5-3B, SmolLM2-1.7B, OLMo-3-7B); for homogeneous families (Gemma-7B, Gemma-2-9B) uniform allocation is already near-optimal.

Experiment 3: structured token-update pruning



Fisher advantage $\Delta CE_{\text{baseline}} - \Delta CE_{\text{Fisher}}$ (symlog; > 0 : κ better).

ΔPPL on WikiText at drop fraction $\rho = 0.10$.

Pruning the least salient fraction $\rho \in \{0.05, 0.1, 0.2, 0.3\}$ of token-indexed updates: $\kappa_{\ell,s}$ is far better than random and gives modest, consistent gains over attention, $J^T J$ trace, diagonal Fisher and grad x act in aggregate.

Experiment 4: predictive-geometry distillation of low-rank students

Student: projections $W \approx AB$, $A \in \mathbb{R}^{d_{\text{out}} \times r}$, $B \in \mathbb{R}^{r \times d_{\text{in}}}$, $r \in \{16, 32, 64\}$; SVD or activation-aware SVD initialisation.

$$\mathcal{L} = \mathcal{L}_{\text{KD}} + \lambda_{\text{CE}} \mathcal{L}_{\text{CE}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}},$$

$$\mathcal{L}_{\text{geo}} = \frac{1}{|S|} \sum_{\ell \in S} \frac{1}{2} \Delta z_{\ell}^{\top} F_T \Delta z_{\ell},$$

$$\Delta z_{\ell} = g_T(h_{\ell}^S) - g_T(h_{\ell}^T),$$

$$g_T = \text{LMHead}_T \circ \text{FinalNorm}_T$$

(readout-Fisher proxy of K_{ℓ}^T). KD objectives: forward KL [10], reverse KL [8], skew KL [13].

KD baseline	Wins	Mean Δ_{Geo}	Median Δ_{Geo}
Forward KL	8/14	-38.4	-12.6
Reverse KL	9/14	+21.3	+12.3
Skew KL	10/14	+59.3	+8.6

$\Delta_{\text{Geo}} = \text{PPL}(\text{KD}) - \text{PPL}(\text{KD}+\text{Geo})$ over 14 matched (model, initialisation, rank, seed) settings; SmoLM2-135M/360M, Qwen2.5-0.5B.

The geometry term complements stronger mode-seeking and skewed KD objectives; it does not improve forward KL on average.

Section 5

Follow-up analyses

Motivation and contributions

Problem setting

Theory

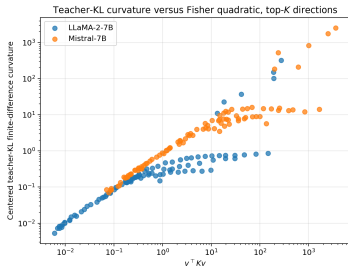
Empirical validation

Follow-up analyses

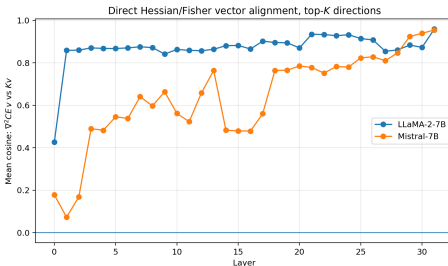
Discussion

Does K_ℓ match the true curvature?

LLaMA-2-7B and Mistral-7B, OpenWebText, all 32 layers, $\rho = 10^{-3}$



Finite-difference teacher-KL curvature vs.
 $v^\top K_\ell v$.

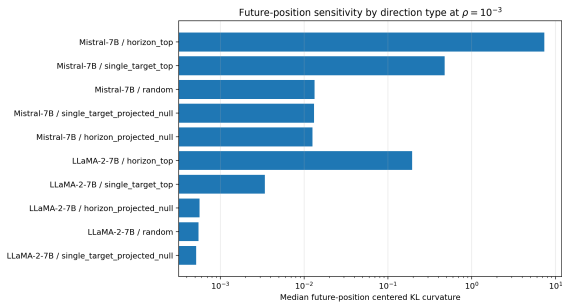


Mean cosine between $\nabla^2 \text{CE} v$ and $K_\ell v$ by layer.

Top-Fisher directions	Spearman($v^\top K_\ell v$, FD)	mean $\cos(H_{\text{CE}} v, K_\ell v)$	median rel. error [†]
LLaMA-2-7B	0.978	0.87	0.20
Mistral-7B	0.975	0.64	0.30

[†] $|v^\top H_{\text{CE}} v - v^\top K_\ell v| / |v^\top H_{\text{CE}} v|$; on null/random directions, where curvature is ≈ 0 , this ratio is 0.94–0.98.

Single-target null directions and future positions



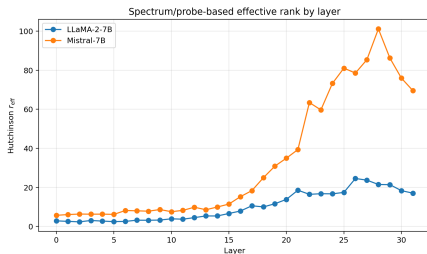
Median future-position KL curvature by direction type,
 $\rho = 10^{-3}$.

Median future-sum FD^2	LLaMA-2-7B	Mistral-7B
Horizon-top	0.195	7.34
Single-target top	0.0034	0.477
Single-target null	0.00052	0.0132
Random	0.00055	0.0133

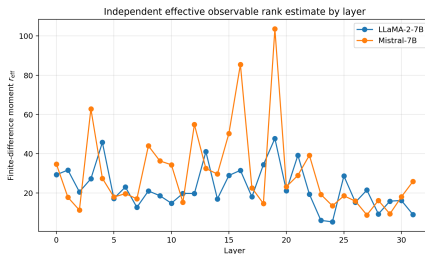
- Null directions of a single target behave like random directions for later positions (future/immediate ratio 3.7 vs. 3.0 and 8.0 vs. 7.9).
- They remain 2–3 orders of magnitude below horizon-aware top directions.

K_ℓ is a per-target object; a horizon-summed operator is the natural multi-target extension.

Effective rank: two estimators



Hutchinson (spectral) r_{eff} by layer.



Finite-difference moment r_{eff} by layer.

	mean FD r_{eff}	mean Hutchinson r_{eff}	Spearman across layers	d
LLaMA-2-7B	22.7	10.1	-0.30	4096
Mistral-7B	30.3	33.2	-0.32	4096

Both estimators give $r_{\text{eff}} = O(10\text{--}100) \ll d$; the layerwise ordering is estimator-dependent, so r_{eff} serves as an order-of-magnitude summary.

Section 6

Discussion

Motivation and contributions

Problem setting

Theory

Empirical validation

Follow-up analyses

Discussion

Limitations and future directions

- **Local by design.** Results hold near a reference trajectory; how local geometries glue together, and how they emerge during training, is open.
- **Perturbation scale.** Small radii approach numerical noise, large radii leave the Taylor regime; adaptive, layer-dependent radii are needed.
- **Architectural scope.** Decoder-only pre-norm causal models; encoder–decoder, bidirectional, MoE and retrieval-augmented models remain.
- **Compute.** Matrix-free, but costlier than activation norms or attention; probe sharing, sketching, amortised estimators.
- **Reduction objectives.** Controlled tests rather than full pipelines; combine Fisher saliency with first-order gradient criteria.
- **Geometry-aware distillation.** Longer budgets, larger students, other low-rank parameterisations, instruction-tuned models.

Conclusion

- Near successful trajectories, the terminal loss induces a pullback Fisher operator $K_\ell = \bar{J}_\ell^\top F^* \bar{J}_\ell$ on each hidden layer, up to $O(\sqrt{\varepsilon_*})$.
- Its spectrum separates output-sensitive from prediction-null directions: a thin, anisotropic tube with $r_{\text{eff}} \ll d$.
- Its token traces $\kappa_{\ell,s}$ are supported on causal ancestors and graded by Jacobian couplings.
- Empirically, the geometry predicts perturbation sensitivity, guides rank allocation and token pruning, and improves low-rank recovery with reverse/skew KL.



Paper repository

References I

- [1] M. Abdin et al. Phi-3 technical report. *arXiv:2404.14219*, 2024.
- [2] L. Ben Allal et al. SmolLM2: When smol goes big. *arXiv:2502.02737*, 2025.
- [3] S. Ashkboos et al. SliceGPT: Compress large language models by deleting rows and columns. *ICLR*, 2024.
- [4] T. Brown et al. Language models are few-shot learners. *NeurIPS*, 2020.
- [5] Gemma Team. Gemma: Open models based on Gemini research and technology. *arXiv:2403.08295*, 2024.
- [6] A. Gokaslan, V. Cohen, E. Pavlick, S. Tellex. OpenWebText corpus, 2019.
- [7] A. Grattafiori et al. The Llama 3 herd of models. *arXiv:2407.21783*, 2024.
- [8] Y. Gu, L. Dong, F. Wei, M. Huang. MiniLLM: Knowledge distillation of large language models. *ICLR*, 2024.
- [9] S. Hayou, N. Ghosh, B. Yu. LoRA+: Efficient low rank adaptation of large models. *ICML*, 2024.
- [10] G. Hinton, O. Vinyals, J. Dean. Distilling the knowledge in a neural network. *NeurIPS Workshop*, 2015.
- [11] E. J. Hu et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 2022.
- [12] A. Q. Jiang et al. Mistral 7B. *arXiv:2310.06825*, 2023.
- [13] J. Ko, S. Kim, T. Chen, S.-Y. Yun. DistiLLM: Towards streamlined distillation for large language models. *ICML*, 2024.
- [14] J. Li et al. Constraint-aware and ranking-distilled token pruning for efficient transformer inference. *KDD*, 2023.
- [15] S. Merity, C. Xiong, J. Bradbury, R. Socher. Pointer sentinel mixture models. *arXiv:1609.07843*, 2016.
- [16] OLMo Team. OLMo 3. *arXiv:2512.13961*, 2025.

References II

- [17] G. Penedo et al. The FineWeb datasets. *arXiv:2406.17557*, 2024.
- [18] J. Tian et al. FLAT-LLM: Fine-grained low-rank activation space transformation. *Findings of EACL*, 2026.
- [19] H. Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.
- [20] L. Valeriani, D. Doimo et al. The geometry of hidden representations of large transformer models. *NeurIPS*, 2023.
- [21] A. Vaswani et al. Attention is all you need. *NeurIPS*, 2017.
- [22] A. Yang et al. Qwen2.5 technical report. *arXiv:2412.15115*, 2025.
- [23] Z. Yuan et al. ASVD: Activation-aware singular value decomposition for compressing large language models. *arXiv:2312.05821*, 2023.
- [24] J. Yun, M. Kim, Y. Kim. Focus on the core: Efficient attention via pruned token compression. *Findings of EMNLP*, 2023.
- [25] H. Wang, B. Dedhia, N. K. Jha. Zero-TPrune: Zero-shot token pruning through leveraging of the attention graph. *arXiv:2305.17328*, 2023.