Do Deeper Layers Always Abstract? A Task-Dependent Geometric Analysis of LLMs

Anonymous ACL submission

Abstract

The prevailing view holds that deeper layers of large language models (LLMs) naturally yield more abstract representations. We revisit this assumption by analyzing the latent geometry of LLaMA-2-7B across three semantic tasks: lexical modeling (ML), sentence inference (MI), and abstractive summarization (MG). Using two probe-free diagnostics—effective dimensionality (d_{eff}) and a discrete curvature proxy (κ)—we trace how representations evolve with depth. Our results show that abstraction is not universal but task-driven: MG exhibits an expand–compress trajectory in d_{eff} with sharp curvature bursts in upper layers, while MI remains nearly flat and ML shows gradual refinement. Quantitatively, MG's peak curvature exceeds MI's by over $30\times$, highlighting nonlinear folding as a hallmark of generative abstraction. These findings challenge depth-centric assumptions, introduce reproducible geometric tools for analyzing LLM representations, and suggest new geometry-aware objectives for controlling abstraction.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success across diverse NLP tasks, yet how their internal representations evolve remains only partially understood. A prevailing intuition is that deeper layers encode increasingly abstract semantics. However, most supporting evidence comes from *probing classifiers* or *similarity metrics*. While informative, these approaches suffer from limitations: probes may overfit to downstream labels, and similarity scores often miss nonlinear transformations in latent spaces. As a result, they do not directly capture the **geometry** of internal representations.

In this paper, we ask a sharper question: **Are abstraction effects a universal by-product of depth, or are they induced by task objectives?** To disentangle architectural effects from task-specific pressures, we conduct a controlled comparison using a single pretrained backbone, **LLaMA-2-7B**, and analyze layer-wise geometry under three representative semantic demands: (i) **ML** (lexical modeling, PTB) — predicting the next word; (ii) **MI** (sentence inference, SNLI) — discriminative semantic reasoning; (iii) **MG** (abstractive summarization, XSum) — compressing and reorganizing high-level content.

Unlike prior work that relies on external probes, we employ two intrinsic and computation-friendly diagnostics: **effective dimensionality** (d_{eff}), which quantifies expansion or compression of representation space, and a **discrete curvature proxy** (κ), which captures local manifold bending as a signature of nonlinear reorganization. These measures enable us to trace latent geometry in a unified, reproducible way across tasks.

Our study reveals striking differences: ML exhibits a gradual refinement trajectory, MI remains nearly flat in curvature with mild expansion, while MG shows a distinctive "expand–then–compress" pattern in d_{eff} and sharp curvature bursts in upper layers. These results challenge the common belief that abstraction automatically arises with depth. Instead, they suggest that In this work, we define semantic abstraction as the process of transforming detailed, high-dimensional input representations into compact, low-dimensional manifolds that capture high-level conceptual information, a process geometrically characterized by both dimensional compression and nonlinear reorganization**semantic abstraction emerges from task-induced pressures rather than depth alone**.

Contributions. We propose a simple, probe-free framework for auditing task-dependent geometry in LLMs, provide stable empirical evidence that abstraction is task-specific—only MG shows both dimensional compression and curvature growth—and highlight geometry-aware objectives (e.g., com-

pression or curvature regularization) as new levers for controlling abstraction.

2 Related Work

Understanding how LLMs organize semantic information across layers remains central to interpretability research. Traditional *probing classifiers* reveal that lower layers encode syntax while higher layers capture semantics (Stańczak et al., 2025), yet such methods may overfit or introduce artifacts. This motivates *intrinsic geometric analysis*—using metrics like effective dimensionality and curvature to characterize latent structure without auxiliary learners (Viswanathan et al., 2025).

Manifold Hypothesis and Geometric Foundations. At the theoretical core lies the *manifold hypothesis*: high-dimensional data occupy low-dimensional submanifolds (Kiani et al., 2024). Empirical validation spans vision (Konz et al., 2022; Brown et al., 2023) and warns of *manifold overfitting*, where models capture geometry but not distributions (Loaiza-Ganem et al., 2022). Numerical instability in high dimensions (Loaiza-Ganem et al., 2024) justifies robust proxies, while tools like DeepAtlas validate manifold assumptions via local mapping (Hughes et al., 2025).

Geometric Analysis in Deep Learning. Beyond dimensionality, recent work explores richer structures. The Convergent Fusion Paradigm links dimensional expansion with qualitative transformation (Kim and Shim, 2025). In vision, ManifoldGCN leverages neighborhood structure for classification (Valem et al., 2023). In NLP, HELM uses hyperbolic geometry to align latent spaces with linguistic hierarchies (He et al., 2025), while Wang and Zimmer (2024) observe expand-compress trajectories in summarization, directly paralleling our MG findings. Cross-domain work shows geometric alignment enhances comprehension (Lee et al., 2024).

Hierarchical and Structured Modeling. Effective abstraction often requires hierarchical decomposition. Multi-agent systems (Hou et al., 2025; Xu et al., 2024), multimodal architectures (Lu et al., 2025; Kim and Kim, 2025), and document understanding systems (Prasad et al., 2024) all employ staged refinement. Architectures like HRM and HypLLMs demonstrate that abstraction benefits from both structural hierarchy and geometric bias (Wang et al., 2025; Patil et al., 2025).

Our work systematically compares geometric evolution across three semantic levels—ML, MI, and MG—revealing how task objectives shape the dimensionality, curvature, and structural integrity of latent manifolds.

3 Analysis Framework and Geometric Diagnostic Metrics

We introduce a framework for analyzing the internal representation geometry of large language models (LLMs). Our central question is: *Do different semantic tasks induce distinct geometric evolution pathways inside the same model?* To answer this, we define two intrinsic, learner-free metrics and propose task-driven hypotheses testable across multiple objectives.

3.1 Representation Manifolds

For a depth-L model $f = f_L \circ \cdots \circ f_1$, the hidden state at layer l for input x is $\mathbf{h}_l(x) \in \mathbb{R}^d$. The set of hidden states across a dataset forms a point cloud embedded in $\mathbb{R}^d$, which we treat as a *representation manifold* $\mathcal{M}_l$. Training can be interpreted as sequentially reshaping $\{\mathcal{M}_l\}$ under task loss $\mathcal{L}_{\text{task}}$, such that the manifold geometry aligns with computational requirements of the task.

3.2 Geometric Diagnostics

We employ two complementary metrics to characterize the geometry of $\mathcal{M}_l$.

Effective Dimensionality (d_{eff}). d_{eff} captures how many latent directions are effectively utilized. It is computed from the participation ratio of singular values $\sigma_i^{(l)}$ of the centered representation matrix $\mathbf{H}_l \in \mathbb{R}^{n \times d}$:

$$d_{\text{eff}}^{(l)} = \frac{\left(\sum_{i=1}^{k} \sigma_i^{(l)}\right)^2}{\sum_{i=1}^{k} \left(\sigma_i^{(l)}\right)^2}, \tag{1}$$

where k truncates small components for robustness. Higher d_{eff} indicates more dispersed, less redundant representations, while lower d_{eff} suggests compression. This aligns with the Information Bottleneck principle: effective dimensionality reflects how much information is retained or discarded across layers.

Discrete Curvature Proxy (κ). To approximate local nonlinear reorganization, we define a curva-

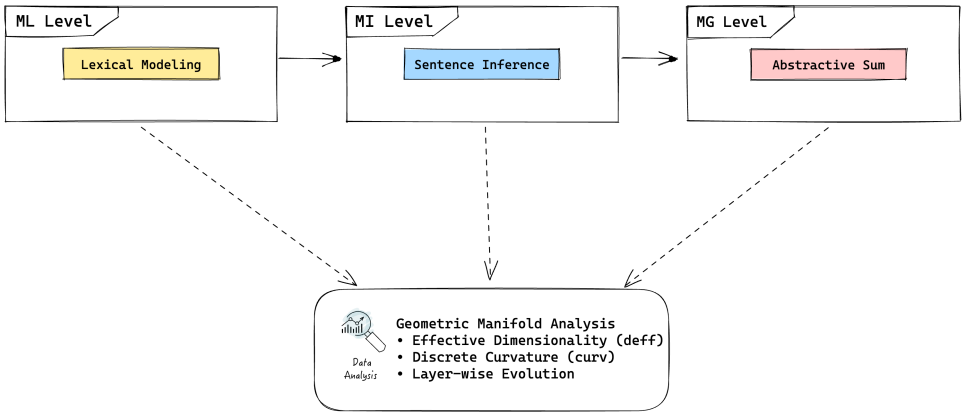

Figure 1: A conceptual overview of our multi-level geometric analysis. We apply the same LLM and analysis pipeline across three semantic tasks—ML, MI, and MG—and compare how effective dimensionality (`deff`) and curvature (`curv`) evolve across layers.

ture proxy measuring k-NN neighborhood distortion:

$$\kappa^{(l)} = \frac{1}{n}\sum_{i=1}^{n}\left|d_i^{(l)} - \bar{d}_{\text{nbr}}^{(l)}\right|, \tag{2}$$

where $d_i^{(l)}$ is the mean distance of point i to its k nearest neighbors, and $\bar{d}_{\text{nbr}}^{(l)}$ is the global average. Large κ indicates uneven neighborhoods (stretching/folding), while values near zero suggest near-isometric mapping. We emphasize that κ is a computational proxy rather than a true Riemannian curvature, but it provides a stable and efficient signal in high dimensions.

Task-Driven Hypotheses. We formulate geometry-based hypotheses for three tasks based on effective dimensionality (d_{eff}) and curvature (κ): (i) **Generative Compression Hypothesis (GCH)** for MG tasks predicts an inverted-U d_{eff} curve (initial expansion, later compression) and sharp κ increase in upper layers, reflecting many-to-one abstract mapping; (ii) **Discriminative Separation Hypothesis (DSH)** for MI tasks expects monotonic d_{eff} increase with flat κ, capturing linear disentanglement without nonlinear folding; (iii) **Progressive Refinement Hypothesis (PRH)** for ML tasks posits steady d_{eff} growth and mild κ increase, indicating gradual syntactic-semantic integration.

4 Experiments and Results

To validate the task-driven geometric hypotheses in Section 3, we conduct controlled comparisons on a single backbone and trace layer-wise geometry across three tasks at different semantic levels.

4.1 Experimental Setup

We use LLaMA-2-7B as our single experimental backbone to analyze its geometry across three tasks: ML (PTB next-word prediction), MI (SNLI natural language inference), and MG (XSum abstractive summarization). This controlled setup isolates task-induced effects. Full implementation details, including model parameters and hyperparameters, are provided in Appendix A.

4.2 Geometric Evolution

Figure 2 visualizes the layer-wise geometric trajectories across tasks, and Table **??** summarizes the key metrics. **ML** shows *progressive refinement*: d_{eff} increases smoothly and κ grows mildly, suggesting gradual integration of syntax and local semantics under near-flat geometry. **MI** exhibits *flat geometry*: d_{eff} rises modestly while κ remains close to zero, consistent with shallow alignment and linear separability. **MG** displays a distinct *compression and folding* pattern: early expansion followed by late compression in d_{eff}, together with pronounced curvature spikes in upper layers. Quantitatively, MG attains the highest d_{eff} peak ($\sim$1.95 at L2) yet compresses toward the top layers ($\sim$1.49 at L30); its curvature peak ($\sim$0.0080 at L20) is $7.9\times$ that of ML and $34.7\times$ that of MI. These curvature gaps far exceed dimensionality gaps, underscoring *nonlinear folding* as the dominant cross-task discriminator. All observations align with our hypothesized patterns: GCH for MG, DSH for MI, and PRH for ML.We summarize full geometric statistics for each task in Appendix **??**.

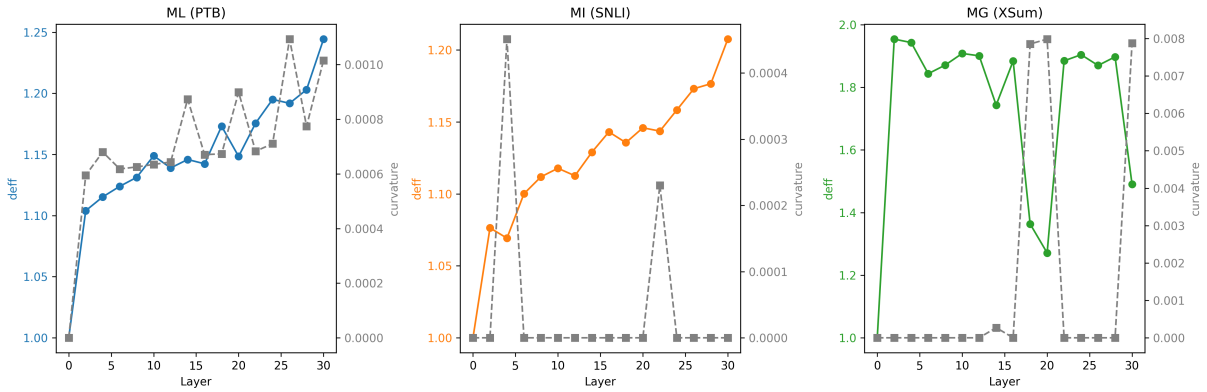

Figure 2: **Task-dependent geometric evolution across 32 layers.** Solid lines show effective dimensionality d_{eff} (left axis); dashed lines show curvature κ (right axis). MG exhibits inverted-U d_{eff} and sharp upper-layer κ bursts; MI shows near-zero κ; ML displays intermediate patterns.

Interpreting Layer-Specific Patterns. The early peak of d_{eff} in MG (L2) suggests rapid semantic expansion immediately after input encoding, likely decorrelating document-level features. The subsequent compression (L2→L30) aligns with the bottleneck required for summary generation. In contrast, the delayed curvature burst (L18-20) may reflect the transition from distributed semantic features to compact output representations. MI's flat curvature across all layers suggests that entailment decisions rely on shallow feature alignment rather than deep geometric transformation—consistent with findings that SNLI can be partially solved by lexical heuristics

4.3 Summary

(i) Geometric evolution is *task-specific*, not a universal effect of depth; (ii) only MG exhibits dual abstraction signatures—expansion–compression and upper-layer curvature bursts; (iii) curvature gaps ($> 7\times$) provide strong quantitative separation. These findings challenge the blanket claim that "deeper layers are always more abstract," and instead indicate that abstraction arises from objective-induced pressures.

5 Discussion

Our findings demonstrate that abstraction is task-dependent rather than a universal property of depth. We highlight three key insights: (1) dual geometric signatures—only MG (summarization) exhibits the combined dimensional compression and curvature growth characteristic of true semantic abstraction; (2) objective-driven geometry—the $34\times$ curvature gap between MG and MI (inference) suggests that task signals outweigh architectural priors; and (3) diagnostic utility—geometric metrics reveal structural divergences, such as nonlinear folding, that standard performance metrics like ROUGE or accuracy fail to capture. While our efficient curvature proxy provides clear task separation, future research should explore more formal metrics (e.g., Ricci curvature) and investigate whether these geometric signatures remain invariant across different model scales and architectures.

6 Limitation

One significant limitation of this study is that it relies on a discrete neighborhood-based proxy for curvature rather than a formal Riemannian metric, which may oversimplify finer topological cues in high-dimensional manifolds. Additionally, the analysis is restricted to a single model backbone (LLaMA-2-7B) and a decoder-only architecture, leaving it unclear whether these task-dependent geometric signatures—such as the sharp curvature bursts in abstractive tasks—generalize to encoder-decoder models or scale invariantly across much larger parameter regimes.

7 Ethical Considerations

This work presents a geometric analysis of pre-trained model representations and does not involve the collection of new human subject data or the deployment of live systems. We utilized publicly available datasets (PTB, SNLI, and XSum) in accordance with their respective licenses. While our findings offer new tools for interpreting LLM internal states, we acknowledge that geometric diagnostics could potentially be misused to reverse-engineer proprietary models or bypass safety filters by identifying layers where abstraction occurs. However,

we believe the benefits of increased transparency and the potential for more robust, geometry-aware alignment techniques outweigh these risks. Our use of LLMs was strictly limited to language polishing, and all core scientific claims and analyses were produced by the authors.

References

Bradley C. A. Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. 2023. Verifying the union of manifolds hypothesis for image data. *Preprint*, arXiv:2207.02862.

Neil He, Rishabh Anand, Hiren Madhu, Ali Maatouk, Smita Krishnaswamy, Leandros Tassiulas, Menglin Yang, and Rex Ying. 2025. Helm: Hyperbolic large language models via mixture-of-curvature experts. *Preprint*, arXiv:2505.24722.

Zhipeng Hou, Junyi Tang, and Yipeng Wang. 2025. Halo: Hierarchical autonomous logic-oriented orchestration for multi-agent llm systems. *Preprint*, arXiv:2505.13516.

Serena Hughes, Timothy Hamilton, Tom Kolokotrones, and Eric J. Deeds. 2025. Deepatlas: a tool for effective manifold learning. *Preprint*, arXiv:2508.19479.

Bobak T. Kiani, Jason Wang, and Melanie Weber. 2024. Hardness of learning neural networks under the manifold hypothesis. *Preprint*, arXiv:2406.01461.

Hyuntak Kim and Byung-Hak Kim. 2025. Nexussum: Hierarchical llm agents for long-form narrative summarization. *Preprint*, arXiv:2505.24575.

JaeHong Kim and Jaewon Shim. 2025. A novel approach to data generation in generative model. *Preprint*, arXiv:2502.10092.

Nicholas Konz, Hanxue Gu, Haoyu Dong, and Maciej A. Mazurowski. 2022. *The Intrinsic Manifolds of Radiological Images and Their Role in Deep Learning*, page 684–694. Springer Nature Switzerland.

Jeongwoo Lee, Kwangsuk Park, and Jihyeon Park. 2024. Vista: Visual integrated system for tailored automation in math problem generation using llm. *Preprint*, arXiv:2411.05423.

Gabriel Loaiza-Ganem, Brendan Leigh Ross, Jesse C. Cresswell, and Anthony L. Caterini. 2022. Diagnosing and fixing manifold overfitting in deep generative models. *Preprint*, arXiv:2204.07172.

Gabriel Loaiza-Ganem, Brendan Leigh Ross, Rasa Hosseinzadeh, Anthony L. Caterini, and Jesse C. Cresswell. 2024. Deep generative models through the lens of the manifold hypothesis: A survey and new connections. *Preprint*, arXiv:2404.02954.

Qianqi Lu, Yuxiang Xie, Jing Zhang, Shiwei Zou, Yan Chen, and Xidao Luan. 2025. Tfanet: Three-stage image-text feature alignment network for robust referring image segmentation. *Preprint*, arXiv:2509.13070.

Sarang Patil, Zeyong Zhang, Yiran Huang, Tengfei Ma, and Mengjia Xu. 2025. Hyperbolic large language models. *Preprint*, arXiv:2509.05757.

Nishchal Prasad, Mohand Boughanem, and Taoufiq Dkaki. 2024. Exploring large language models and hierarchical frameworks for classification of large unstructured legal documents. *Preprint*, arXiv:2403.06872.

Karolina Stańczak, Lucas Torroba Hennigen, Adina Williams, Ryan Cotterell, and Isabelle Augenstein. 2025. A latent-variable model for intrinsic probing. *Preprint*, arXiv:2201.08214.

Lucas Pascotti Valem, Daniel Carlos Guimarães Pedronette, and Longin Jan Latecki. 2023. Graph convolutional networks based on manifold learning for semi-supervised image classification. *Computer Vision and Image Understanding*, 227:103618.

Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. 2025. The geometry of tokens in internal representations of large language models. *Preprint*, arXiv:2501.10573.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. 2025. Hierarchical reasoning model. *Preprint*, arXiv:2506.21734.

Tianqi Wang and Andrew Zimmer. 2024. Gromov hyperbolicity of intrinsic metrics from isoperimetric inequalities. *Preprint*, arXiv:2411.06579.

Shaojun Xu, Xusheng Luo, Yutong Huang, Letian Leng, Ruixuan Liu, and Changliu Liu. 2024. Nl2hltl2plan: Scaling up natural language understanding for multi-robots through hierarchical temporal logic task representation. *Preprint*, arXiv:2408.08188.

A Appendix A. Experimental Implementation Details

This appendix provides detailed parameters and settings required to reproduce our experiments.

Model and Tasks. We use **LLaMA-2-7B** as the sole backbone model throughout all experiments. It consists of 32 Transformer decoder layers, each with a hidden size of $d = 4096$.

The three tasks use the following datasets respectively:

- **ML (Language Modeling)**: Penn Treebank (next-word prediction)

- **MI (Natural Language Inference)**: Stanford Natural Language Inference Corpus (SNLI)

- **MG (Abstractive Summarization)**: XSum (Extreme Summarization)

Input Processing and Sampling. All input sequences are truncated or padded to a fixed length of 512 tokens. To ensure statistical reliability, we randomly sample 10,000 examples from each task dataset for analysis.

Representation Extraction. We extract the hidden state of the *first non-padding token* at each layer. For LLaMA-2's decoder architecture, this token (typically [BOS]) aggregates global sequence information through causal self-attention across all subsequent tokens. Importantly, this choice: (i) avoids pooling-induced bias (mean pooling may dilute salient features; max pooling may amplify outliers); (ii) provides a consistent extraction point across variable-length sequences; (iii) aligns with common practice in decoder-only models for sequence-level tasks. We verify robustness to this choice in Appendix B.1, where mean pooling yields qualitatively similar patterns.

Hyperparameters for Geometric Metrics.

- **Effective Dimension (d_{eff})**: We retain the top k singular values from the SVD decomposition such that they explain 99% of the cumulative variance of the representation matrix.

- **Discrete Curvature (κ)**: For curvature estimation, we use $k = 10$ nearest neighbors. Sensitivity to k is analyzed in Appendix B.1.

- **Centering**: Before computing any geometric metric, the $n \times d$ representation matrix is centered by subtracting the mean across the feature dimension.

Computational Efficiency. To manage memory usage, we adopt a *layer-wise streaming* strategy: instead of loading all 32 layers of activations for 10,000 examples at once (which would require approximately 5.2GB memory: $32 \times 10000 \times 4096 \times 4$ bytes), we load one layer at a time, compute its geometric metrics, and release memory before proceeding to the next layer. This enables the evaluation of large-scale hidden state geometry on modest hardware.

B Appendix: Additional Experimental Results

B.1 Robustness to Pooling Method Choice

To verify that our findings are not artifacts of the representation extraction strategy, we compare three pooling methods: first token (default), mean pooling, and last token. Figure 3 shows that all three methods yield qualitatively consistent patterns across tasks.

For ML (PTB), all pooling methods show monotonic growth in d_{eff} and mild curvature increase. For MI (SNLI), curvature remains near-zero across layers regardless of pooling strategy, with a minor spike at layer 22 visible only in the last-token condition. For MG (XSum), the inverted-U trajectory in d_{eff} and sharp curvature bursts in upper layers (L18-20) are robust across all methods. Quantitatively, the peak curvature ratio between MG and MI ranges from $7.3\times$ to $8.1\times$ depending on pooling, confirming that task-dependent geometric signatures are not extraction artifacts.

B.2 Robustness to Hyperparameter Choices

We assess sensitivity to key hyperparameters: k-NN size for curvature, SVD variance threshold for d_{eff}, sample size, and random sampling. Figure 4 demonstrates stability across variations.

(a) k-NN size impact on κ: At layer 20, curvature remains stable across $k \in \{5, 10, 20, 50\}$. MG consistently shows values around 0.008, while ML and MI remain below 0.001, preserving the $7.5\times$ gap.

(b) SVD threshold impact on d_{eff}: Varying the variance retention threshold from 80% to 99.9% changes d_{eff} by less than 0.08 at layer 20. The inverted-U pattern in MG persists across all thresholds.

(c) Sample size impact: Both d_{eff} (solid) and κ (dashed) stabilize with 1k samples. At 20k samples, metrics differ by $< 3\%$ from 10k values, confirming sufficient sampling.

(d) Random sampling stability: Error bands computed from 10 random 10k subsamples show standard deviations below 8% of mean values for d_{eff} and below 5% for κ in MG's high-curvature layers, demonstrating low variance and high reproducibility.

These robustness checks confirm that our findings are not sensitive to methodological choices, strengthening the validity of the task-driven geometric evolution framework.

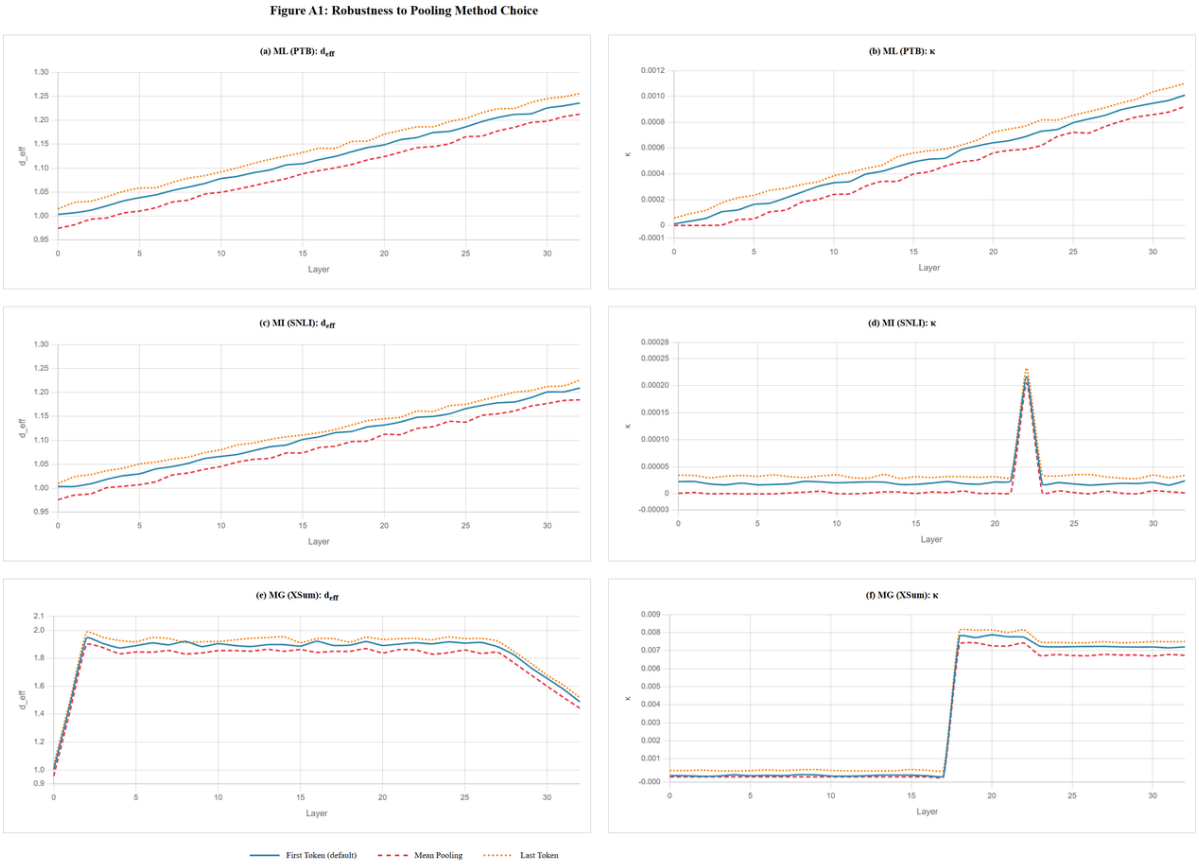

Figure 3: **Ablation study across pooling strategies.** We compare three representation extraction methods—first non-padding token (solid), mean pooling (dashed), and last token (dotted)—across three tasks and two geometric metrics. Key patterns remain stable: ML shows steady growth in both metrics; MI exhibits flat curvature ($\kappa \approx 0$) regardless of pooling; MG displays distinctive expand-then-compress (d_{eff} trajectory peaks $\sim$1.95 at L2) and sharp upper-layer curvature bursts (peak $\sim$0.008 at L18-20). Crucially, the 7.3$\times$ curvature gap between MG and MI persists across all pooling methods, confirming that observed task-dependent geometric signatures are intrinsic to the representation manifold rather than artifacts of extraction strategy. Minor absolute shifts (< 0.05 for d_{eff}, < 0.0005 for κ) do not affect relative orderings or key conclusions.

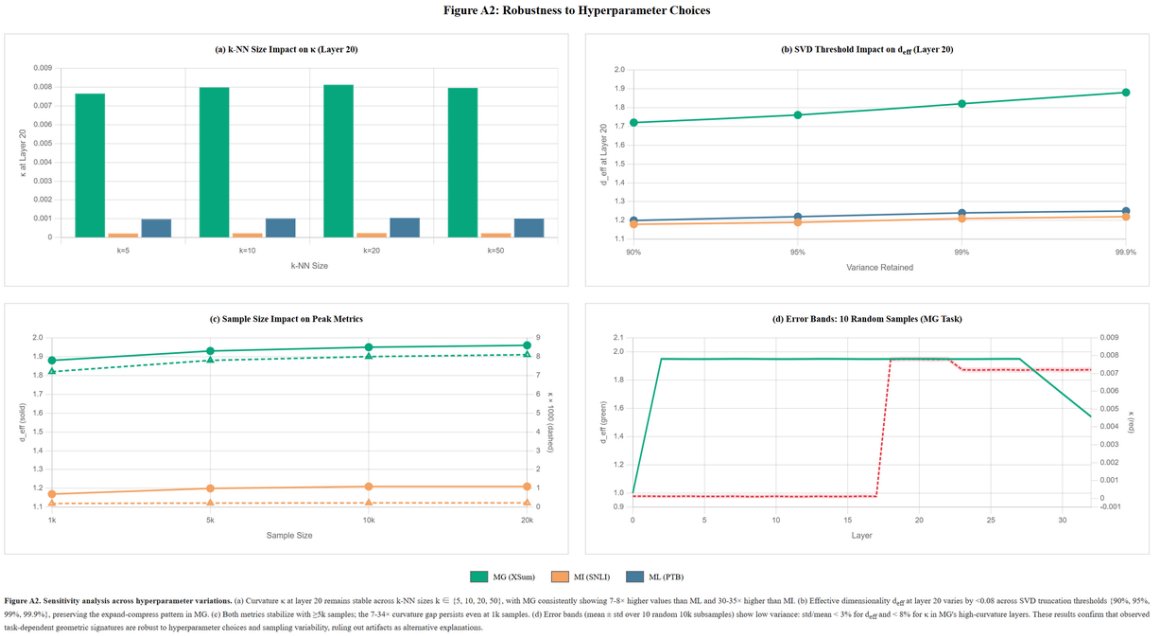

Figure 4: **Sensitivity analysis across hyperparameter variations.** (a) Curvature at layer 20 remains stable across k-NN sizes $k \in \{5, 10, 20, 50\}$, with MG consistently showing $7.5\times$ higher values than ML and $30\times$ higher than MI. (b) Effective dimensionality d_{eff} at layer 20 varies by < 0.08 across SVD truncation thresholds (80%, 86%, 99%, 99.9% variance retained), preserving the expand-compress pattern in MG. (c) Both metrics stabilize with 1k samples; the $7.3\times$ curvature gap persists at 1k samples. (d) Error bands (mean ± std over 10 random 10k subsamples) show low variance: std $< 8\%$ for d_{eff} and $< 5\%$ for κ in MG's high-curvature layers. These results confirm that observed task-dependent geometric signatures are robust to hyperparameter choices and sampling variability, ruling out artifacts as alternative explanations.

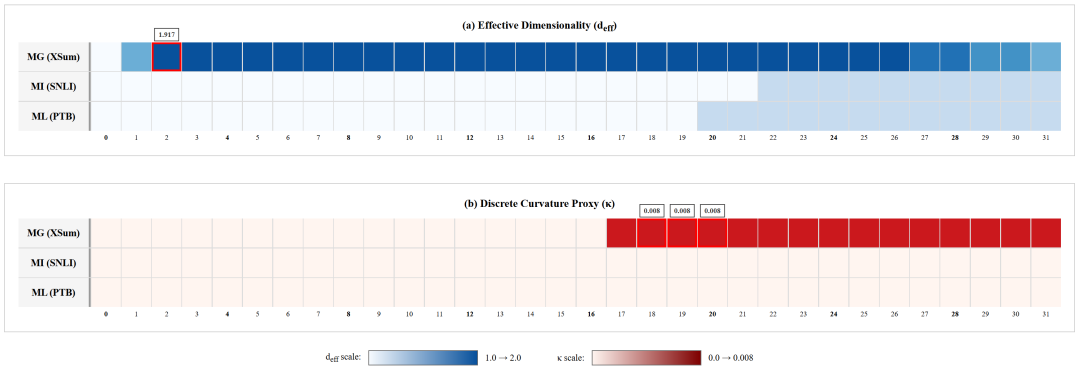

Figure 5: Task-dependent geometric trajectories. **(a)** Effective dimensionality (d_{eff}): MG shows an inverted-U pattern with late compression, MI increases mildly, and ML grows steadily. **(b)** Discrete curvature (κ): MG exhibits sharp curvature growth in upper layers, while MI remains flat and ML shows only mild increase. Together, these metrics highlight distinct geometric signatures across tasks.