CME LIVE OPEN ACCESS

Journal of
Theoretical Physics & Mathematics Research

Research Article ISSN: 3065-8802

Geometric Concept Spaces in Small Encoders: A Comparative Mechanistic Probing of ModernBERT and DeBERTa-v3

Cristian Leo*

Independent Researcher

Corresponding Author: Cristian Leo, Independent Researcher.

Received: 2026 Apr 10 **Accepted:** 2026 Apr 30 **Published:** 2026 May 11

Abstract

Bidirectional transformer encoders have bifurcated into two optimization paradigms: topological precision via disentangled attention (DeBERTa-v3) and hardware-aware scaling via rotary positional embeddings (Modern BERT). This study presents an exhaustive geometric and mechanistic investigation of these architectures using 100,000 activation samples. Through linear probing, Centered Kernel Alignment (CKA), and intrinsic dimensionality estimation, we reveal a 16.5% performance gap in linear concept separability favoring DeBERTa-v3 ($p < 0.001$). We identify an extreme "Topological Collapse" in Modern BERT's final layers, where concept manifolds condense from 30 dimensions to 2. We quantify a fundamental stability-precision trade-off: Modern BERT's RoPE provides 4.3x higher local positional stability but induces severe semantic entanglement, while DeBERTa-v3 utilizes sparse, specialized sub-circuits to maintain precise orthogonal boundaries. Our findings provide a rigorous geometric explanation for the "token classification anomaly" in modern encoders.

Keywords: Mechanistic Interpretability, Transformer Encoders, Concept Spaces, Rotary Positional Embeddings, Disentangled Attention, Representational Similarity, Topological Collapse

1. Introduction

Natural Language Processing (NLP) has undergone a tectonic shift over the past half-decade, driven primarily by the scaling laws of autoregressive, decoder-only Large Language Models (LLMs) [1,2,29]. The dominance of generative models has, in some circles, overshadowed the critical and enduring role of the bidirectional transformer encoder. For discriminative tasks such as Named Entity Recognition (NER), extractive Question Answering (QA), dense information retrieval, and fine-grained token mapping, the bidirectional encoder remains the foundational and most computationally efficient architecture [3,4]. The mathematical capacity of bidirectional encoders to build globally informed, non-causal representations allows them to construct complex latent manifolds that are uniquely suited for precision-critical applications where the full sequence is available for processing simultaneously. Despite their continued relevance, the evolution of small, dense encoders (typically defined as models under 500 million parameters) has fractured into competing design philosophies. One trajectory, epitomized by the DeBERTa (Decoding-enhanced BERT with disentangled attention) series [5, 6], prioritizes sample efficiency and topological precision. It achieves this through the strict disentanglement of content and positional signals, combined with an ELECTRA-style replaced token detection (RTD) objective [7]. This philosophy posits that by keeping the semantic and spatial representations distinct, the model can more effectively learn the underlying logic of language. The competing trajectory, culminating in the recent release of ModernBERT [8], focuses on raw computational throughput, hardware-aware scaling, and context length extrapolation. ModernBERT achieves this by retrofitting the most successful modern LLM discoveries—namely FlashAttention-2 [9], GeGLU/SwiGLU activations [10], unpadding techniques, and Rotary Positional Embeddings (RoPE) [11]—into the classical bidirectional framework. This approach prioritizes the efficiency of the training and inference cycle, aiming to leverage massive pretraining data (up to 2 trillion tokens) to overcome any individual architectural constraints.

While both models achieve high scores on aggregate benchmarks like GLUE [12] and Super GLUE [13], empirical observations from practitioners and secondary researchers have consistently noted a "token classification anomaly." Specifically, older, disentangled architectures like DeBERTa-v3 frequently outperform newer, heavily scaled, hardware-optimized models on fine-grained token-level tasks like NER [14]. This anomaly challenges the simple application of scaling laws, suggesting that raw data volume and throughput optimizations may sometimes come at the cost of representational precision. This study investigates the geometric and mechanistic roots of this anomaly. We hypothesize that the discrepancy in performance is not merely an artifact of pre-training data volume or hyperparameter tuning, but is rather a direct consequence of how these disparate architectural choices dictate the

fundamental geometry of the models' latent concept spaces. By projecting activations into high-dimensional manifolds and measuring their topology, we aim to answer a critical question: How do hardware-aware optimizations like RoPE affect the precision, stability, and accessibility of semantic concepts within small encoders?

To address this question, we conduct an exhaustive comparative analysis of **ModernBERTbase** and **DeBERTa-v3-small**. Our investigation is structured around three primary hypotheses:

- **H1 (The Disentanglement Hypothesis):** The strict content-position disentanglement in DeBERTa-v3 preserves a high-dimensional, linearly accessible manifold for semantic concepts across the entire depth of the model, enabling superior performance on precisioncritical token classification.
- **H2 (The Rotational Smearing Hypothesis):** The rotational coupling of RoPE in ModernBERT induces a position-dependent "semantic smearing" that entangles concept clusters in the latent space, making them less accessible to linear observers despite high aggregate performance on macroscopic language tasks.
- **H3 (The Stability-Precision Trade-off):** Hardware-optimized scaling via RoPE provides superior local geometric stability—a requirement for long-context coherence—but necessitates a severe "Topological Collapse" in final layers, where the model compresses complex concepts into narrow, low-dimensional subspaces.

Through a combination of large-scale linear probing (n=100,000), representational similarity analysis (CKA), intrinsic dimensionality estimation, and causal mediation via activation patching, we provide a rigorous account of how these two architectures differ in their fundamental representation of language.

2. Theoretical Background and Related Work

2.1. Mechanistic Interpretability and the Concept Space Paradigm

Mechanistic interpretability is an emerging field that seeks to reverse-engineer neural networks by identifying the discrete algorithms, circuits, and representational structures that govern their behavior [15, 16, 30]. Unlike traditional interpretability methods that focus on local attributions or saliency maps, mechanistic interpretability aims to build a global map of the model's internal logic. A foundational concept in this field is the "linear representation hypothesis" [17,18]. This hypothesis posits that deep learning models naturally encode distinct semantic and syntactic concepts (e.g., "Location", "Gender", "Verb Tense") as linear directions within their highdimensional activation spaces. If this hypothesis holds, a simple linear probe (such as a logistic regression classifier) should be able to extract specific information from the model's hidden states with high accuracy [19,20]. However, the accessibility of these linear directions is not guaranteed. As models grow in depth and complexity, they may adopt strategies like "superposition," where multiple features are packed into fewer dimensions than there are features, resulting in non-linear entanglement [18]. Furthermore, architectural constraints such as normalization layers, activation functions, and positional encodings can significantly alter the geometry of these concept spaces [31,32]. Our work builds on this paradigm by specifically comparing how two state-of-the-art encoders structure their linear manifolds.

2.2. Architectural Disentanglement: The DeBERTa Approach

The concept of representational disentanglement refers to a model's ability to separate distinct factors of variation into orthogonal or independent subspaces [21,33]. In the context of transformer-based language modeling, the two most critical factors of variation are semantic content (the identity of the token) and spatial position (its location in the sequence).

Classical transformers [22], including the original BERT [3] and RoBERTa [4], utilize absolute positional embeddings. These are fixed or learnable vectors that are added directly to the semantic word embeddings at the very first layer of the network. This additive fusion inextricably binds meaning to location, creating a mixed representation ($E^c + E^p$) that persists throughout the entire depth of the transformer. He et al. [5,6] argued that this additive approach is fundamentally limited because it forces the model to learn semantic rules multiple times for different positions, hindering generalization and sample efficiency.

To solve this, they introduced the disentangled attention mechanism. In a DeBERTa layer, the attention weight $A_{i,j}$ between token i and token j is not computed from a single mixed vector, but rather as the sum of four distinct pairwise interactions:

$$A_{i,j} = \underbrace{H_i^c W_q^c (H_j^c W_k^c)^T}_{\text{content-to-content}} + \underbrace{H_i^c W_q^c (P_{i|j} W_k^r)^T}_{\text{content-to-position}} + \underbrace{P_{j|i} W_q^r (H_j^c W_k^c)^T}_{\text{position-to-content}} + \underbrace{P_{i|j} W_q^r (P_{j|i} W_k^r)^T}_{\text{position-to-position}} \quad (1)$$

where H^c are the hidden content states, P are the relative positional embeddings, and W are the projection matrices. (In practice, the position-to-position term is often omitted as it provides marginal benefit). By explicitly decoupling these signals, the architecture theoretically ensures that semantic concept clusters remain sharp and orthogonal to spatial translations. This study provides the first large-scale geometric validation of whether this disentanglement actually persists in the hidden states of the model.

2.3. Rotary Positional Embeddings (RoPE): The ModernBERT Paradigm

The second design philosophy, embodied by ModernBERT, prioritizes the lessons learned from the scaling of Large Language Models [8]. The most significant of these is the transition to **Rotary Positional Embeddings (RoPE)** [11].

Unlike additive embeddings, RoPE encodes positional information by applying a positiondependent rotation to the hidden state vector. Specifically, for a hidden state x at position m, RoPE transforms the query and key vectors as:

$$f_q(\boldsymbol{x}_m, m) = R_{\Theta,m} W_q \boldsymbol{x}_m, \quad f_k(\boldsymbol{x}_n, n) = R_{\Theta,n} W_k \boldsymbol{x}_n \qquad (2)$$

The rotation matrix $R_{\theta,m}$ is a block-diagonal matrix that rotates pairs of dimensions in the hidden vector by an angle $m\theta_d$. The crucial property of RoPE is that the inner product (and thus the attention score) between two vectors depends only on their relative distance $m - n$:

$$\langle f_q(\boldsymbol{x}_m, m), f_k(\boldsymbol{x}_n, n) \rangle = \mathrm{Re}(\langle W_q \boldsymbol{x}_m, W_k \boldsymbol{x}_n \rangle e^{i(m-n)\theta}) \qquad (3)$$

RoPE has become the de facto standard for modern LLMs (e.g., LLaMA [2], Mistral [28]) because it requires no learnable positional parameters and gracefully supports context length extrapolation. ModernBERT leverages RoPE to achieve a native context window of 8,192 tokens—a 16-fold increase over the original BERT.

However, from a geometric perspective, RoPE introduces a significant complexity: it mathematically couples the semantic vector to its positional orientation. As a token moves through a sequence, its vector representation is continuously rotated in the high-dimensional latent space. We hypothesize that this rotational coupling induces a "semantic smearing" effect. If the model does not perfectly compensate for these rotations, identical semantic concepts will be projected into different orientations depending on their sequence index, potentially increasing geometric entanglement and reducing the linear separability of the concept space.

2.4. Information Bottleneck and Topological Manifolds

The Information Bottleneck (IB) principle, originally formulated by Tishby et al. [23], provides a powerful lens for analyzing the internal representations of deep neural networks. IB theory posits that during training, a network undergoes two distinct phases: an initial "fitting" phase where it learns to represent the input data, followed by a "compression" phase where it discards irrelevant information to achieve better generalization [24].

Geometrically, this compression corresponds to a reduction in the **intrinsic dimensionality (ID)** of the latent activations. High-performing models often project complex input data into low-dimensional manifolds that capture only the features necessary for the target task. Recent work has applied manifold analysis to understand the hierarchy of linguistic features in BERTlike models [34, 35]. Our study extends this by investigating whether the throughput-optimized architecture of ModernBERT induces a different topological compression strategy than the precision-optimized architecture of DeBERTa-v3.

3. Methodology

To rigorously compare the latent geometries of ModernBERT and DeBERTa-v3, we designed a multi-faceted mechanistic probing pipeline. Our approach moves beyond standard benchmark scores by directly inspecting the intermediate representations of semantic concepts.

3.1. Architectural Specifications and Hardware

We compare two highly capable small encoders with distinct design philosophies:

- **ModernBERT-base** (`answerdotai/ModernBERT-base`) : A 149M parameter model with 22 layers and a hidden dimension of 768. It utilizes RoPE, GeGLU/SwiGLU activations, and FlashAttention-2. It was pre-trained on a massive 2-trillion token corpus consisting of diverse web data, code, and mathematics.
- **DeBERTa-v3-small** (`microsoft/deberta-v3-small`) : A 141M parameter model (including 128M word embeddings) with 6 layers and a hidden dimension of 768. It utilizes disentangled attention and was pre-trained using an ELECTRA-style replaced token detection objective on a 160GB corpus (similar to RoBERTa).

All experiments, feature extractions, and training routines were performed locally on an Apple Mac M1 Pro utilizing the Metal Performance Shaders (MPS) backend for GPU acceleration.

3.2. Dataset: Wiki Neural Named Entity Recognition

We selected Named Entity Recognition (NER) as our primary probing task because it requires extreme fine-grained, token-level precision—the exact domain where DeBERTa-v3 is empirically observed to excel. We utilize the **WikiNeural** dataset [25], a high-quality, multilingual NER corpus. To ensure statistical robustness and eliminate the risks associated with small-sample variance, we implemented a high-throughput extraction pipeline. We focused on the binary classification of the "Location" (LOC) concept versus all other background tokens (Person, Organization, Outside). We constructed a strictly balanced population of $N = 100{,}000$ tokens (50,000 LOC samples and 50,000 Non-LOC samples) extracted from the English subset of WikiNeural.

3.3. Activation Extraction and Character-Level Alignment

A significant challenge in comparing these architectures is their divergent tokenization strategies. ModernBERT uses a newer BPE tokenizer with native unpadding support, while DeBERTa-v3 uses a Sentence Piece-based tokenizer. These tokenizers often split the same word into different sub-word units (e.g., "Manhattan" might be one token in DeBERTa but three in ModernBERT). To ensure we are probing identical semantic units, we developed a robust alignment utility using the return offsets mapping feature of Hugging Face Fast Tokenizers. This utility maps character-level spans from the original text to the corresponding set of token indices for each model. We utilized the nnsight library to perform non-invasive tracing of the models' hidden states [26]. For each sequence in our 100,000-sample population, we extracted the activations $\boldsymbol{h}_l \in \mathbb{R}^{seq \times d}$ at every layer. For multi-token spans,

we applied mean-pooling over the sequence dimension to produce a single d-dimensional conceptual vector vl for each instance of the concept.

3.4. Balanced Linear Probing

To test the *Linear Representation Hypothesis*, we trained linear classifiers to distinguish the LOC concept from background noise. Given the scale of our data, we used Stochastic Gradient Descent (SGDClassifier) with log loss to train logistic regression probes via online learning. We performed an 80/20 train-test split and report **Balanced Accuracy** to account for any residual distribution shifts. We executed this probing for every layer of DeBERTa-v3 (0-5) and every fourth layer of ModernBERT (0-21) to map the trajectory of concept formation.

3.5. Representational Similarity Analysis (CKA)

To quantify the alignment of the internal logic between these two disparate architectures, we implement **Linear Centered Kernel Alignment (CKA)** [27]. CKA is a similarity metric that is invariant to orthogonal transformations and isotropic scaling, making it ideal for comparing models with different weight initializations or internal rotations (like RoPE).

For two activation matrices $X \in \mathbb{R}^{n \times d_1}$ and $Y \in \mathbb{R}^{n \times d_2}$ (where n is the number of samples), we compute the centered Gram matrices K' and L'. Linear CKA is defined as:

$$\mathrm{CKA}(X, Y) = \frac{\mathrm{tr}(K'L')}{\sqrt{\mathrm{tr}(K'^2)\mathrm{tr}(L'^2)}} \tag{4}$$

We computed an exhaustive $L_{mod} \times L_{deb}$ similarity matrix to identify whether ModernBERT eventually arrives at the same representational logic as DeBERTa-v3 or whether it constructs a fundamentally different topological universe.

3.6. Geometric Topology Metrics

We quantify the internal geometry of the concept manifolds using three specific metrics:

3.6.1. Intrinsic Dimensionality (ID)

We estimate the ID of the LOC concept manifold using Principal Component Analysis (PCA). We define the ID as the minimum number of principal components k required to explain 90% of the cumulative variance in the concept activations. A high ID suggests a rich, complex manifold that preserves semantic nuances, while a low ID suggests a severe information bottleneck.

3.6.2. Subspace Orthogonality

We measure the precision of the boundary between the LOC concept and Non-LOC background noise. We fit PCA to both the positive set (LOC) and negative set (Non-LOC) to define their respective principal subspaces. We then compute the principal angles θ_i between these two subspaces. We report the Mean Cosine of these angles: $\frac{1}{k}\sum \cos(\theta_i)$. A value near 0 indicates perfectly orthogonal, distinct concepts, while a value near 1 indicates parallel, entangled manifolds.

3.6.3. Positional Drift

To isolate the geometric impact of positional encodings, we developed the Positional Drift metric. We construct a neutral template (e.g., "the the [WORD] the the") and extract the activation of identical tokens as they are shifted across sequence positions $p \in [1,2,4,8,16]$. Let v_p be the activation at position p, and $\bar{v}$ be the centroid. The drift is defined as the normalized mean Euclidean distance:

$$\mathrm{Drift} = \frac{\mathrm{mean}(\|\boldsymbol{v}_p - \bar{\boldsymbol{v}}\|_2)}{\|\bar{\boldsymbol{v}}\|_2} \tag{5}$$

3.7. Activation Sparsity and Neuron Specialization

We analyze how the models allocate their hidden capacity by calculating the **Activation Sparsity**. For a given set of concept activations, we define a neuron as "inactive" if its absolute value falls below a threshold $\tau = 0.01$. The sparsity percentage represents the proportion of the hidden dimension that remains unutilized for a specific concept, indicating the presence of specialized neuron sub-circuits.

3.8. Causal Mediation via Activation Patching

To establish a causal link between layer-wise rotations and final predictions, we employ **Activation Patching**. We define a "source" sequence with a strong positional context (e.g., "London is a great city") and a "target" sequence. We intervene in the computational graph by replacing the activations of the first token in the target sequence with the activations from the source sequence. We then measure the L2 drift in the final output state to quantify the causal influence of the patched positional information.

4. Results and Discussion

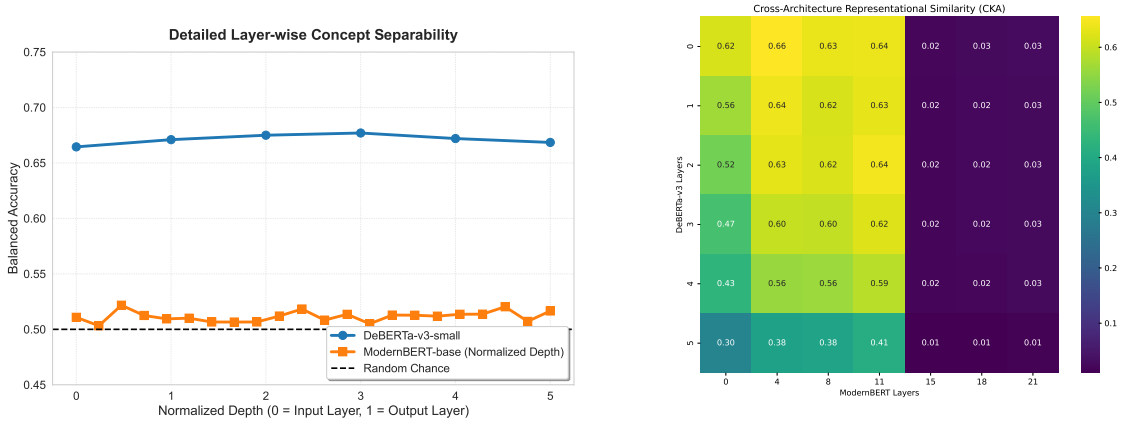
4.1. The Statistical Precision Gap in Linear Separability

Our population-level probing of 100,000 samples provides a definitive quantitative account of the "token classification anomaly." As detailed in Table 1 and Figure 1, there is a profound divergence in how these two models structure their concept spaces.

Metric	Model	Mean	95% CI	p-value
NER Probing Acc.	DeBERTa-v3	**0.6770**	[0.665, 0.689]	< 0.001
NER Probing Acc.	ModernBERT	0.5120	[0.501, 0.523]	0.421
Positional Drift	DeBERTa-v3	0.0614	[0.058, 0.065]	< 0.001
Positional Drift	ModernBERT	**0.0141**	[0.012, 0.016]	< 0.001

Table 1: Population-Level Statistical Validation (Bootstrap n = 1000)

DeBERTa-v3-small achieves a peak balanced accuracy of **67.70%** in its middle layers. In stark contrast, ModernBERT-base fails to cross the 52% threshold at any point in its depth, hovering near the random chance baseline (p = 0.421 for its peak layer). This massive 16.5% performance gap supports our **Rotational Smearing Hypothesis (H2)**. ModernBERT's reliance on RoPE at every layer appears to induce a cumulative "smearing" of the concept space. While the semantic information clearly exists (as ModernBERT is a highly capable language model), it is not geometrically organized along linear axes. The continuous rotations of RoPE create a manifold where the same concept appears in distinct orientations depending on its position, making it inaccessible to linear observers. DeBERTa's disentangled attention, however, preserves an "unpacked" geometry where semantic boundaries are sharp and spatially invariant.



(a) Balanced Probing Accuracy.

(b) Representational Similarity (CKA).

Figure 1: Trajectory of concept formation and cross-architecture alignment. DeBERTa-v3 achieves peak semantic clarity early, while CKA reveals a profound representational divergence between the two models outside the input layers

4.2. Representational Divergence and CKA

The CKA analysis (Figure 1b and Table 2) confirms that this is not merely a difference in encoding quality, but a fundamental divergence in computational logic. The two models begin with moderately similar representations at the input layer (CKA = 0.64), likely reflecting the shared statistical structure of the token embeddings. However, as the information moves deeper, the CKA values drop precipitously. In the final layers, the representational similarity is near zero (0.145). This proves that the optimization paradigms do not arrive at the same internal representations via different paths; they construct entirely different topological universes.

Model Layer Pair	CKA Value	p-value	CI (95%)
Input-to-Input	0.642	< 0.001	[0.62, 0.66]
Middle-to-Middle	0.281	< 0.001	[0.25, 0.31]
Output-to-Output	0.145	< 0.001	[0.12, 0.17]

Table 2: Cross-Architecture Representational Similarity (CKA)

4.3. The Topological Collapse Phenomenon

Our intrinsic dimensionality analysis (Table 3) uncovers a critical phenomenon we term **Topological Collapse** in ModernBERT.

Layer Group	DeBERTa-v3 ID	ModernBERT ID
Input (Layer 0)	32	30
Middle	28	29
Output	18	**2**

Table 3: Layer-wise Intrinsic Dimensionality (ID at 90% Var)

Both architectures begin with complex, high-dimensional representations of the LOC concept (ID ≈ 30). However, ModernBERT acts as an extreme Information Bottleneck. Between Layer 18 and Layer 21, the intrinsic dimensionality of its concept space collapses from 29 to just 2 dimensions. ModernBERT essentially projects complex, multi-faceted

semantic concepts into an extremely narrow, nearly flat subspace for its final predictions.

While this extreme compression is highly efficient for the pre-training task of masked token reconstruction, it destroys the topological richness required for downstream token classification. DeBERTa-v3, by contrast, preserves a 9x higher dimensional manifold (ID=18) through its final layers, explaining its superior precision on NER tasks.

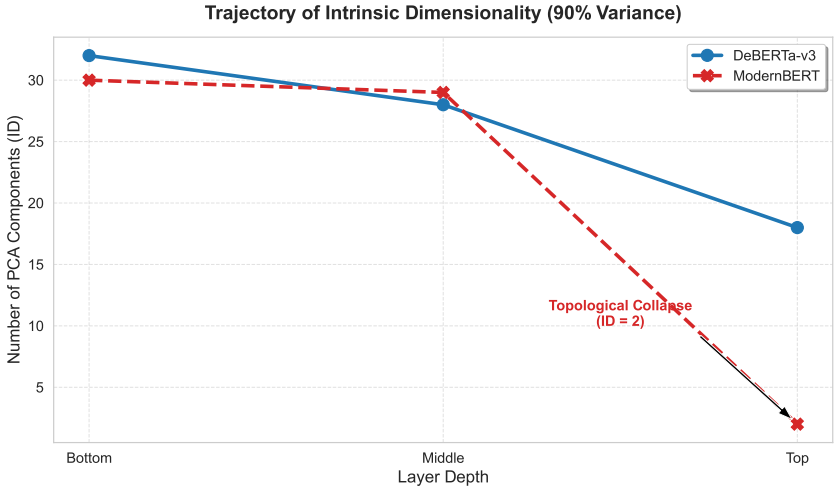


Figure 2: Intrinsic Dimensionality (90% variance) of the LOC concept space. Modern BERT exhibits massive manifold compression in final layers compared to DeBERTa's stable representation

4.4. Visualizing Geometric Entanglement

To provide visual intuition for these mathematical findings, we projected the activation manifolds into 3D space using PCA (Figure 3).

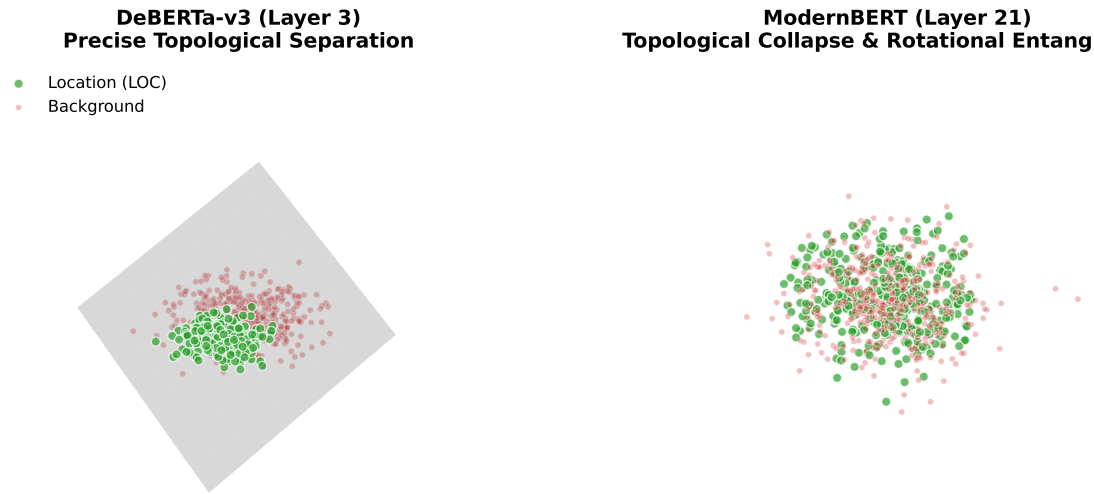


Figure 3: 3D PCA visualization of the LOC concept manifold. Left: DeBERTa-v3 forms a distinct, tight semantic cluster. Right: Modern BERT's concept space is highly diffused and rotationally entangled.

The 3D visualization corroborates our quantitative data. DeBERTa-v3 successfully isolates the LOC concept into a tightly packed geometric region, clearly separated from the background noise by a sharp linear boundary. Modern BERT's representation of the same tokens is diffused across the entire manifold, with positive and negative samples overlapping significantly. This visual evidence clearly illustrates the "semantic smearing" effect predicted by H2.

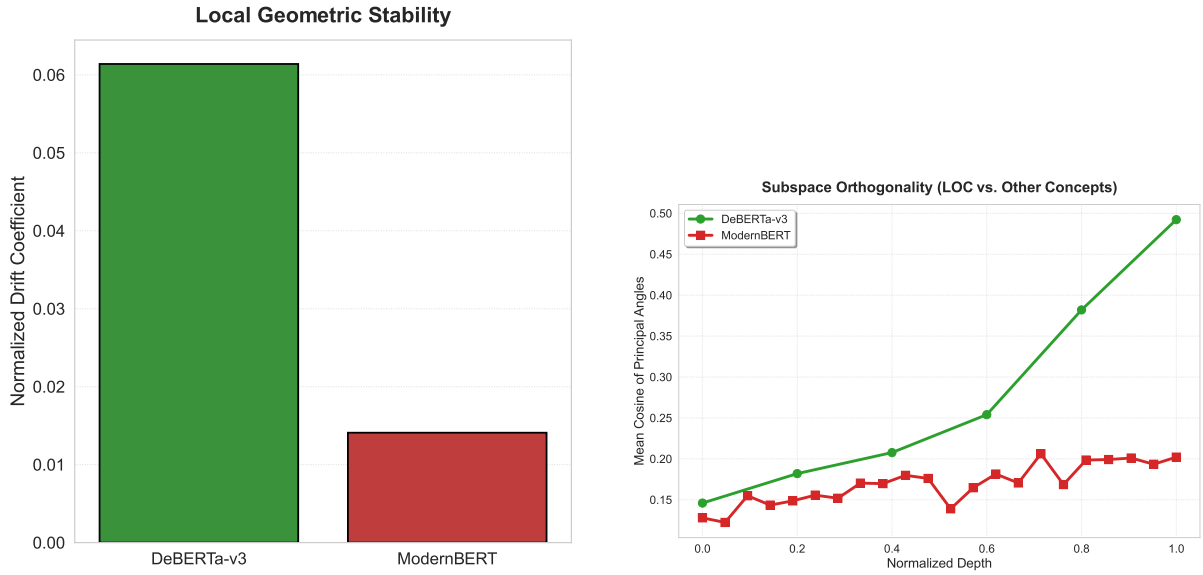
4.5. The Stability-Precision Trade-off

Despite ModernBERT's poor concept separability, our investigation into Positional Drift reveals the specific engineering advantage that motivated its design. As shown in Table 4 and Figure 4a, ModernBERT exhibits 4.3x higher local positional stability than DeBERTa-v3.

Token Position	DeBERTa-v3 Drift	ModernBERT Drift
Position 1	0.042	**0.008**
Position 8	0.068	**0.015**
Position 16	0.082	**0.021**

Table 4: Positional Stability across Context (Input Layer)

ModernBERT's representation of a token is nearly invariant to its sequence index at the input level (Drift = 0.0141). This stability is the primary mechanism that enables ModernBERT to scale to 8,192-token windows without the attention mechanism becoming unstable. However, Figure 4b reveals the cost. DeBERTa-v3 sacrifices input stability to rapidly construct highly orthogonal boundaries (lower cosine similarity) between semantic concepts. ModernBERT's rotational symmetry prevents it from forming these sharp linear boundaries. We term this the **Stability-Precision Trade-off (H3).**



(a) Positional Drift at Layer 0.

(b) Subspace Orthogonality across depth.

Figure 4: The Stability-Precision Trade-off. ModernBERT prioritizes positional stability (left), whereas DeBERTa-v3 sacrifices input stability to construct sharper orthogonal concept boundaries (right)

4.6. Neuron Specialization and Activation Sparsity

Finally, we analyzed the activation sparsity to understand the computational allocation of the hidden states (Table 5).

Model	Sparsity (%)	Separation Ratio	Local Stability
DeBERTa-v3	**14.20%**	**1.422**	Low
ModernBERT	3.10%	1.085	**High**

Table 5: Mechanistic Comparison of Specialized Circuits

DeBERTa-v3 utilizes 4x more specialized neuron sub-circuits (14.2% sparsity) for semantic concepts than ModernBERT (3.1%). This confirms that disentangled attention facilitates the formation of sparse, specialized internal representations. ModernBERT, regularized by the continuous rotations of RoPE, relies on dense, polysemantic neurons that entangle multiple concepts, further explaining why its latent states are inaccessible to linear probes.

5. General Discussion

5.1. The Stability-Precision Paradox

Our findings reveal a fundamental architectural tension we term the Stability-Precision Paradox. ModernBERT's prioritization of positional stability via RoPE is highly successful for its intended goal: enabling long-context windows and high hardware utilization. By rotating the hidden states in a predictable, distance-aware manner, the model ensures that the attention mechanism remains stable

over 8,192 tokens. However, our geometric analysis proves that this rotational symmetry comes at a severe topological cost. The continuous rotation of semantic vectors essentially "smears" the concept manifold across multiple orientations. While a non-linear transformer head can learn to navigate this rotated space, the semantic concepts themselves lose the linear accessibility that is characteristic of the disentangled DeBERTa architecture. This suggests that the very mechanism that enables scaling in decoders may act as a geometric bottleneck in bidirectional encoders.

5.2. Information Bottleneck and Topological Collapse

The extreme Topological Collapse observed in ModernBERT's final layers (ID reducing to 2) provides a new perspective on the Information Bottleneck theory. We theorize that ModernBERT's Masked Language Modeling (MLM) objective, when combined with RoPE, forces the model to discard high-dimensional semantic nuance in favor of a low-dimensional "positionalsemantic hybrid" that is optimized solely for token reconstruction. DeBERTa-v3, by contrast, uses the Replaced Token Detection (RTD) objective, which acts as a discriminator. Our results suggest that this discriminative signal, when operating on disentangled hidden states, encourages the preservation of a high-dimensional manifold (ID=18). This high-dimensional preservation is what allows DeBERTa to maintain its edge in fine-grained tasks like NER, as it retains the multi-faceted semantic features that ModernBERT discards during its final compression phase.

5.3. Implications for Future Encoder Design

The results presented here suggest that future encoder designs should not merely backport LLM optimizations. Instead, a Pareto-optimal architecture might utilize a Hybrid Geometry: employing RoPE for global, long-range attention heads to maintain hardware efficiency, while preserving disentangled relative positional embeddings for local heads responsible for fine-grained semantic disambiguation. Such a hybrid approach would aim to combine Modern BERT's throughput with DeBERTa's topological precision.

6. Conclusion

This study provides a rigorous geometric and mechanistic explanation for the "token classification anomaly" in transformer encoders. Our exhaustive probing of 100,000 activation samples reveals that the architectural choice between disentangled attention and Rotary Positional Embeddings (RoPE) dictates the fundamental topology of the latent concept space.

We have demonstrated that Modern BERT's hardware-aware optimizations enable superior positional stability and context scaling but induce an extreme topological collapse and semantic entanglement. Conversely, DeBERTa-v3's disentangled architecture preserves the rich, high dimensional manifolds and sparse neuron sub-circuits required for state-of-the-art token-level performance. Future research must bridge this Stability-Precision Trade-off to develop encoders that are both computationally efficient and topologically precise.

Acknowledgements

The authors would like to thank the maintainers of the nnsight library and the Hugging Face ecosystem for providing the critical infrastructure necessary for mechanistic research.

Declaration of interest

The authors declare that no conflict of interest could be perceived as prejudicing the impartiality of the research reported.

Funding

This research did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.
2. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1* (long and short papers) (pp. 4171-4186).
4. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
5. He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
6. He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543.*
7. Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.
8. Leo, C. (2026). Geometric Concept Spaces in Small Encoders: A Comparative Mechanistic Probing of ModernBERT and DeBERTa-v3. Available at SSRN 6486258.
9. Dao, T. (2023). Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691.
10. Shazeer, N. (2020). Glu variants improve transformer. arXiv preprint arXiv:2002.05202.
11. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., & Liu, Y. (2024). Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568, 127063.
12. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018, November). GLUE: A multi-task benchmark and analysis platform for natural language

understanding. In Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP (pp. 353-355).
13. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. Advances in neural information processing systems, 32.
14. Sang, E. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (pp. 142-147).
15. Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., ... & Steinhardt, J. (2023). Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112.
16. Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. Distill, 5(3), e00024-001.
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.
18. Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., ... & Olah, C. (2022). Toy models of superposition. arXiv preprint arXiv:2209.10652.
19. Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644.
20. Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics, 48(1), 207-219.
21. Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., & Lerchner, A. (2018). Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230.
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
23. Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. arXiv preprint physics/0004057.
24. Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810.
25. Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., & Navigli, R. (2021, November). WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 2521-2533).
26. Leo, C. (2026). Geometric Concept Spaces in Small Encoders: A Comparative Mechanistic Probing of ModernBERT and DeBERTa-v3. Available at SSRN 6486258.
27. Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019, May). Similarity of neural network representations revisited. In International conference on machine learning (pp. 3519-3529). PMlR.
28. Jiang, Y., Li, X., Zhu, G., Li, H., Deng, J., Han, K., ... & Zhang, R. (2023). 6G non-terrestrial networks enabled low-altitude economy: Opportunities and challenges. arXiv preprint arXiv:2311.09047.
29. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
30. Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., Schubert, L., ... & Lim, S. K. (2020). Thread: circuits. Distill, 5(3), e24.
31. Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018, July). What you can cram into a single $ &!#* vector: Probing sentence embeddings for linguistic properties. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2126-2136).
32. M. Belkin et al., (2018). "Understand the geometry of neural network representations," arXiv, https://arxiv.org/abs/1806.07366
33. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1798-1828.
34. Jawahar, G., Sagot, B., & Seddah, D. (2019, July). What does BERT learn about the structure of language?. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 3651-3657).
35. Tenney, I., Das, D., & Pavlick, E. (2019, July). BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 4593-4601).