The Dimensionality Illusion: Why PCA Variance Does Not Equal Semantic Information in Text Embeddings

Joel Thorarinson* Allison Hensgen†

May 2026

Abstract

Recent work claims that text embeddings with nominal dimensionality of 768–4096 have effective intrinsic dimensionality of approximately 16, as measured by PCA variance explained and downstream task performance on standard benchmarks. This conclusion is an artifact of evaluation methodology, not a property of the embeddings themselves. Using domain-specific vectors spanning legal documents, conversational text, and structured records, we demonstrate that fine-grained intra-domain semantic structure degrades sharply when embeddings are reduced to 16 dimensions via PCA, even when coarse inter-domain classification is preserved. In a 62-term legal dictionary experiment, intra-domain term pairs collapse by up to $\Delta = +0.140$ in cosine similarity, while inter-domain pairs never collapse — explaining why MTEB-style benchmarks report "no degradation." We introduce three complementary evaluation tools: domain-conditional retrieval precision (DCRP), Semantic Coherence Loss (SCL), and Compression-Induced Semantic Aliasing (CISA). An information-theoretic argument establishes that PCA variance is a poor proxy for semantic information: dimensions with low variance can carry high mutual information with specialized query distributions. Claims of low intrinsic dimensionality confuse the geometry of embedding spaces with their information content. Compression must be evaluated against the semantic resolution required by the domain.

Keywords: text embeddings; PCA; dimensionality reduction; retrieval; semantic coherence; semantic aliasing; domain-specific retrieval; DCRP; RAG

1 Introduction

Dense text embeddings now underpin retrieval, classification, and clustering across virtually every NLP application. Since Word2Vec [Mikolov et al., 2013] and GloVe [Pennington et al., 2014], the dimensionality of learned representations has steadily increased: from 300-dimensional word vectors through BERT's 768-dimensional contextual embeddings [Devlin et al., 2019] to modern models such as E5-Mistral [Wang et al., 2024], BGE [Xiao et al., 2023], and nomic-embed [Nussbaum et al., 2024], which produce embeddings in $\mathbb{R}^d$ where $d \in \{768, 1024, 4096\}$, built on the transformer architecture [Vaswani et al., 2017]. A natural question follows: do these embeddings truly utilize all d dimensions, or is the effective dimensionality much lower?

Recent work argues for the latter. Tsukagoshi and Sasano [2025] find that E5-Mistral's 4096-dimensional embeddings can be reduced to 16 dimensions via PCA with "almost no performance degradation" on standard benchmarks. Takeshita et al. [2025b] show that randomly removing 50% of dimensions has "minimal impact" on retrieval and classification tasks. The implication is that text embeddings are massively redundant.

*Coherence Research Group. ORCID: 0000-0002-0553-842X. joel.thorarinson@conformalmaps.com

†Coherence Research Group. ORCID: 0009-0008-7247-0307. AH contributed SCL, CISA, and coherence threshold framework.

This conclusion is an illusion created by the evaluation methodology. Standard benchmarks — MTEB [Muennighoff et al., 2023] for general embeddings and BEIR [Thakur et al., 2021] for zero-shot retrieval — measure coarse-grained discrimination: can the system distinguish documents about law from documents about medicine? This inter-domain discrimination lives in the principal components — the 16 directions of maximum variance. But real-world retrieval systems, particularly those deployed in specialized domains such as law [Chalkidis et al., 2020, 2022, Butler et al., 2025], must perform *intra-domain* discrimination: can the system distinguish a document about the German *Pflichtteil* from one about the French *réserve héréditaire*? This fine-grained semantic structure lives in the long tail of dimensions that PCA discards.

1.1 Contributions

This paper makes the following contributions:

1. We identify **two distinct failure modes** of PCA compression — collapse and distortion — and show that collapse is invisible to existing benchmarks while distortion amplifies category separation, producing the false appearance of robustness.

2. We demonstrate in a 62-term legal dictionary experiment that intra-domain term pairs collapse by up to $\Delta = +0.140$ in cosine similarity at 16 dimensions, while inter-domain pairs never collapse ($\max \Delta = -0.023$), directly explaining why MTEB reports "no degradation."

3. We introduce **Semantic Coherence Loss** (SCL) and **Compression-Induced Semantic Aliasing** (CISA) as diagnostic metrics that detect the failure modes standard benchmarks miss.

4. We provide an information-theoretic proof that PCA variance is a necessary but not sufficient condition for semantic utility: low-variance dimensions can carry arbitrarily high mutual information with domain-specific query distributions.

5. We propose **domain-conditional retrieval precision** (DCRP) and **neighborhood-rank preservation** as evaluation tools for validating embedding compression against the semantic resolution a domain requires.

2 Background and Related Work

2.1 Intrinsic Dimensionality of Embeddings

Intrinsic dimensionality (ID) — the minimum number of parameters needed to represent data without loss of essential structure — has a long history in manifold learning [Tenenbaum et al., 2000, Roweis and Saul, 2000]. For word embeddings, Yin and Shen [2018] showed that performance peaks at specific dimensionalities and degrades beyond them, suggesting a finite ID. TwoNN estimators [Facco et al., 2017] applied to token embeddings yield IDs around 10 dimensions, with increasing redundancy as model scale grows [Kataiwa et al., 2025]. Ruppik et al. [2025] refine this picture by showing that local intrinsic dimension varies across data regions — a finding that should caution against global claims about "the" intrinsic dimensionality of an embedding space.

Most relevant to our argument, Robinson et al. [2025] demonstrate that token embeddings *violate the manifold hypothesis*: neighborhoods lack well-defined intrinsic dimension, and the distribution of local dimensions rejects the assumption of a globally smooth manifold. This undermines the theoretical basis for PCA-based compression, which assumes linear subspace structure that these embeddings provably lack.

2.2 Dimensionality Reduction for Embeddings

PCA has been the primary tool for embedding compression. Mu and Viswanath [2018] showed that removing the top principal components ("all-but-the-top" post-processing) improves word embedding isotropy and downstream performance — already hinting at the complex relationship between variance and information. Raunak [2017] proposed simple PCA-based reduction, and Raunak et al. [2019] extended this with a PPA–PCA–PPA pipeline that halves dimensionality while maintaining word similarity scores. Tsukagoshi and Sasano [2025] applied these ideas to prompt-based embeddings, finding that the first 25% of dimensions carry most variance. Takeshita et al. [2025b] push further, showing that *random* removal of 50% of dimensions has minimal impact — a result more consistent with benchmarks too coarse to detect information loss than with the information being truly absent. Zhang et al. [2024] evaluate multiple unsupervised methods and find PCA dominant; Bengtsson [2025] apply PCA to LLM weight matrices with similar claims of lossless compression.

The random projection literature, beginning with the Johnson–Lindenstrauss lemma [Johnson and Lindenstrauss, 1984] and its efficient variants [Achlioptas, 2003], provides theoretical underpinning: pairwise distances can be approximately preserved in logarithmically many dimensions. However, the JL lemma guarantees pairwise *distance* preservation, not semantic *structure* preservation — a distinction our work makes precise.

The Matryoshka Representation Learning (MRL) framework [Kusupati et al., 2022] trains embeddings to be useful at multiple dimensionalities by design, with post-hoc adaptations [Yoon et al., 2024] and training refinements [Zhang et al., 2025]. Takeshita et al. [2025a] provide a key nuance: embeddings are robust to truncation *except in heavy truncation scenarios* — precisely the regime (16 dimensions) where the strongest claims are made. For RAG deployment, Huerga-Perez et al. [2025] show that PCA + float8 quantization achieves 8× compression with <0.3% loss on MTEB — but MTEB is the evaluator, and the question is whether it detects the right failures.

2.3 Theoretical Limitations of Dense Retrieval

Two recent theoretical results strengthen our argument. Weller et al. [2025] prove that embedding dimension fundamentally limits the number of distinct top-k result sets a retrieval system can return: for any fixed d, there exist document collections where no dot-product embedding can represent certain valid query results. Killingback et al. [2026] establish scaling laws linking embedding dimension to retrieval performance as task complexity increases — performance does *not* plateau at 16 dimensions for complex tasks.

These results provide the theoretical floor for our empirical findings: if dimension limits representational capacity (which it provably does), then discarding dimensions via PCA must discard representational capacity. The question is whether existing benchmarks detect it. We show they do not.

2.4 The Benchmark Problem

MTEB [Muennighoff et al., 2023] covers retrieval, classification, clustering, pair classification, reranking, and STS across multiple datasets. BEIR [Thakur et al., 2021] evaluates zero-shot retrieval across 18 diverse datasets. Both inherit evaluation paradigms from earlier dense retrieval work [Karpukhin et al., 2020, Reimers and Gurevych, 2019], and both are predominantly *inter-domain*: they measure whether a system can distinguish broad categories. Few tasks require fine-grained *intra-domain* discrimination within a specialized corpus. Domain-specific benchmarks such as LexGLUE [Chalkidis et al., 2022] and MLEB [Butler et al., 2025] have begun to address this gap for legal NLP, but dimensionality reduction claims are not typically validated against them. When Tsukagoshi and Sasano [2025] report "almost no degradation" at 16 dimensions, they report that the system can still tell law from medicine — not that it can distinguish a

German *Pflichtteil* from a French *réserve héréditaire*.

3 Theoretical Framework

3.1 PCA Variance ≠ Semantic Information

Let $\mathbf{x} \in \mathbb{R}^d$ be an embedding vector and q be a query. The semantic utility of dimension i is not its variance $\mathrm{Var}(x_i)$ but its mutual information with the relevance label:

$$I(x_i; \mathrm{rel}(q)) = H(\mathrm{rel}(q)) - H(\mathrm{rel}(q) \mid x_i) \tag{1}$$

PCA selects dimensions maximizing $\sum_i \mathrm{Var}(x_i)$, which is equivalent to maximizing total variance. But mutual information [Shannon, 1948, Cover and Thomas, 2006] and variance are fundamentally different quantities:

Proposition 1. *There exist embedding spaces where a dimension x_j has arbitrarily low variance but arbitrarily high mutual information with the relevance labels for a specific query distribution Q.*

Proof. Consider a domain-specific feature that activates (takes value 1) for exactly ϵ-fraction of documents (those in a specialized domain) and is 0 otherwise. Its variance is $\epsilon(1-\epsilon) \approx \epsilon$ for small ϵ. But for queries targeting that domain, this dimension perfectly discriminates relevant from irrelevant documents, giving $I(x_j; \mathrm{rel}(q)) = H(\mathrm{rel}(q))$ — the maximum possible mutual information. □

The dictionary collapse experiment (Section 5.1) provides empirical evidence that analogous sparse features exist in real embedding spaces: the forced heirship terms that collapse under PCA are precisely the low-variance, domain-specific features this proposition predicts.

3.2 The Long Tail of Semantic Information

Text embedding dimensions partition naturally into two regimes:

1. **Principal regime** (dimensions 1–k, $k \ll d$): High variance, encode coarse semantic categories (topic, language, register). Captured by PCA. Sufficient for inter-domain discrimination.

2. **Long-tail regime** (dimensions $k+1$–d): Low variance individually, but collectively encode fine-grained domain-specific semantics. Not captured by PCA. Required for intra-domain retrieval.

The information lost by discarding the long tail is:

$$I_{\text{lost}} = I(\mathbf{x}; \mathrm{rel}(q)) - I(\mathbf{x}_{1:k}; \mathrm{rel}(q)) \tag{2}$$

By the data processing inequality, $I_{\text{lost}} \geq 0$: projecting to the first k PCA components can only discard information. The key insight is that I_{lost} can be large even when $\sum_{i=k+1}^{d} \mathrm{Var}(x_i)$ is small, because variance and mutual information are fundamentally different quantities (Proposition 1).

3.3 Semantic Coherence Loss (SCL)

If PCA compression preserves variance but destroys meaning, we need a metric that measures meaning directly. The relevant question is not how much variance survives compression, but whether the similarity relationships that matter to the domain remain intact. A legal retrieval system that reverses the relative ordering of "coercive contempt" and "punitive contempt" has failed, regardless of how much variance it retains.

We define Semantic Coherence Loss as the degree to which compression disrupts local meaning-order. For a set of domain concepts, let S_{full} be the pairwise similarity structure in the full embedding space, and let S_k be the corresponding structure after reduction to k dimensions:

$$\text{SCL}_k = 1 - \rho(S_{\text{full}}, S_k) \tag{3}$$

where ρ is the Spearman rank correlation over all pairwise similarities. $\text{SCL}_k = 0$ when neighborhood structure is perfectly preserved; it increases as local meaning-order breaks down. SCL complements DCRP: DCRP measures retrieval success at the task level; SCL measures whether the underlying semantic field remains ordered.

3.4 Compression-Induced Semantic Aliasing (CISA)

Beyond reordering neighborhoods, aggressive compression can cause a more severe failure: semantically distinct concepts map to the same point in the reduced space. We call this *Compression-Induced Semantic Aliasing* (CISA), borrowing from signal processing where aliasing occurs when distinct signals become indistinguishable after undersampling.

Formally, a semantic aliasing event occurs when a pair of concepts (a, b) known to be operationally distinct increases in cosine similarity by more than a collapse threshold δ: the pair is *aliased* at dimensionality k if $\text{sim}_k(a, b) - \text{sim}_{\text{full}}(a, b) > \delta$. We use $\delta = 0.1$ throughout.

CISA is dangerous precisely because it is silent. A retrieval system may still return documents from the correct domain while failing to distinguish the correct doctrine, jurisdiction, or compliance requirement. In broad benchmarks, aliased concepts remain inside the same coarse category and therefore escape detection.

4 Experimental Setup

4.1 Corpus

We use three domain-specific vector collections totaling over 2 million points:

Collection	Points	Dims	Domain	Model
Legal documents	244,000	768	Emails, filings, financial	nomic-embed-text
Case documents	1,700,000	768	Full legal corpus	BGE + SPLADE++
Conversational	19,000	768	Messaging, contacts	nomic-embed-text

Table 1: Vector collections used in experiments.

4.2 Query Sets

We construct two types of query sets:

Inter-domain queries (100 queries): Queries that discriminate between broad domains ("find legal documents" vs. "find personal messages"). These are the type of queries well-served by MTEB-style benchmarks.

Intra-domain queries (200 queries): Queries requiring fine-grained discrimination *within* the legal domain. Examples:

- "German forced heirship Pflichtteil" vs. "French réserve héréditaire" (distinguishing specific legal concepts)
- "child support contempt ability to pay" vs. "child support contempt willful refusal" (distinguishing legal standards)
- "Turner v. Rogers procedural safeguards" vs. "general civil contempt procedure" (distinguishing case-specific doctrine)

4.3 Dimensionality Reduction

We apply PCA to reduce embeddings from 768 to $k \in \{16, 32, 64, 128, 256, 384, 512, 768\}$ dimensions. For each reduction level, we measure:

1. **Recall@10**: Standard retrieval metric (MTEB-style)
2. **DCRP@10**: Domain-conditional retrieval precision — Recall@10 computed only for intra-domain queries against same-domain documents
3. **SCL$_k$**: Semantic Coherence Loss — Spearman rank decorrelation of pairwise similarities
4. **NP$_k$**: Neighborhood preservation — Jaccard overlap of top-m neighbors at full vs. reduced dimensionality
5. **CISA count**: Number of semantic aliasing events (distinct pairs crossing the collapse threshold)
6. **Variance explained**: Cumulative variance ratio from PCA

4.4 Neighborhood Preservation

For each concept or query q, let $N_{\text{full}}(q, m)$ be the top-m nearest neighbors in the full embedding space, and $N_k(q, m)$ be the top-m nearest neighbors after PCA reduction to k dimensions. Neighborhood preservation is measured by Jaccard overlap:

$$\text{NP}_k(q, m) = \frac{|N_{\text{full}}(q, m) \cap N_k(q, m)|}{|N_{\text{full}}(q, m) \cup N_k(q, m)|} \tag{4}$$

A sharp decline in NP indicates that the reduced embedding has reordered local semantic topology even when broad cluster membership survives.

4.5 Domain-Conditional Retrieval Precision (DCRP)

For a query q in domain $\mathcal{D}$, let $\text{rel}(q) \subset \mathcal{D}$ be the set of relevant documents (all within the same domain). DCRP@k is:

$$\text{DCRP@}k = \frac{1}{|Q_{\mathcal{D}}|} \sum_{q \in Q_{\mathcal{D}}} \frac{|\text{top-}k(q) \cap \text{rel}(q)|}{\min(k, |\text{rel}(q)|)} \tag{5}$$

where top-$k(q)$ are the k nearest neighbors to q in the (reduced) embedding space, and $Q_{\mathcal{D}}$ are queries targeting domain $\mathcal{D}$. The denominator normalizes by the smaller of k and the number of relevant documents, preventing penalization when fewer than k relevant documents exist.

DCRP measures what domain-specific retrieval requires: finding the *right* documents within a specialized corpus, not just documents from the right broad category.

5 Results

5.1 Experiment 1: Dictionary Collapse

We embedded 62 legal terms from Black's Law Dictionary and cross-jurisdictional legal terminology using nomic-embed-text (768 dimensions), then applied PCA compression to 16 and 32 dimensions. For each term pair, we measured cosine similarity at full dimensionality and after compression. Table 2 shows the most significant results.

PCA to 16 dimensions explained only 63.5% of variance for our 62-term set — substantially less than the 85–95% reported by Tsukagoshi and Sasano [2025] for E5-Mistral on MTEB-style embeddings. At corpus scale, the gap is even larger: PCA at 16 dimensions explains only 38.9% of variance for our 244K-document legal collection and 36.2% for the 1.7M-document case corpus (Table 1). When tested on the full Bouvier's Law Dictionary (6,200 terms), PCA at 16 dimensions captures only 34% of variance. This consistent 35–40% range for domain-specific corpora versus 85–95% for general benchmarks suggests a fundamental difference in variance structure, not a sample-size artifact.

Term A	**Term B**	**768d**	**16d**	**Δ**	**Operational Risk**
CISA: Compression-Induced Semantic Aliasing (distinct terms become indistinguishable)					
coercive contempt	punitive contempt	0.784	0.913	+0.129	Jury trial right lost
legitieme portie	*yuryubun*	0.763	0.890	+0.127	Jurisdiction distinctions blurred
legitieme portie	*iryubun*	0.757	0.884	+0.127	Cross-system comparison fails
nondischargeable	insolvency	0.703	0.814	+0.112	Debt permanence confused
legitieme portie	*laglott*	0.762	0.872	+0.110	Tradition-specific law lost
joint custody	sole custody	0.836	0.932	+0.096	Opposite outcomes merge
Distortion: related terms torn apart					
faraid	intestate succession	0.554	0.026	−0.528	Equivalent concepts unlinked
Sorgerecht	*gemensam vårdnad*	0.640	0.184	−0.456	Parental care concepts separated
domestic support obl.	discharge	0.655	0.288	−0.367	DSO exception invisible
default	contumacy	0.569	0.202	−0.367	Willfulness distinction lost
ability to pay	willful refusal	0.638	0.317	−0.322	Turner safeguard distinction fails
Pflichtteil	elective share	0.709	0.418	−0.291	Children vs. spouse protection confused
Control: inter-domain pairs (never collapse)					
medical malpractice	custody	0.560	−0.168	−0.728	N/A — correctly separated
patent infringement	forced heirship	0.565	0.168	−0.397	N/A — correctly separated

Table 2: Dictionary collapse at 16 dimensions with operational risk assessment. **Top**: CISA events — intra-domain pairs that alias under compression, destroying legally meaningful distinctions. **Middle**: distortion events — related concepts torn apart. **Bottom**: inter-domain controls, which never alias. Model: nomic-embed-text v1.5, 768d. $n = 62$ terms.

Figure 1: Two Failure Modes of PCA Compression to 16 Dimensions

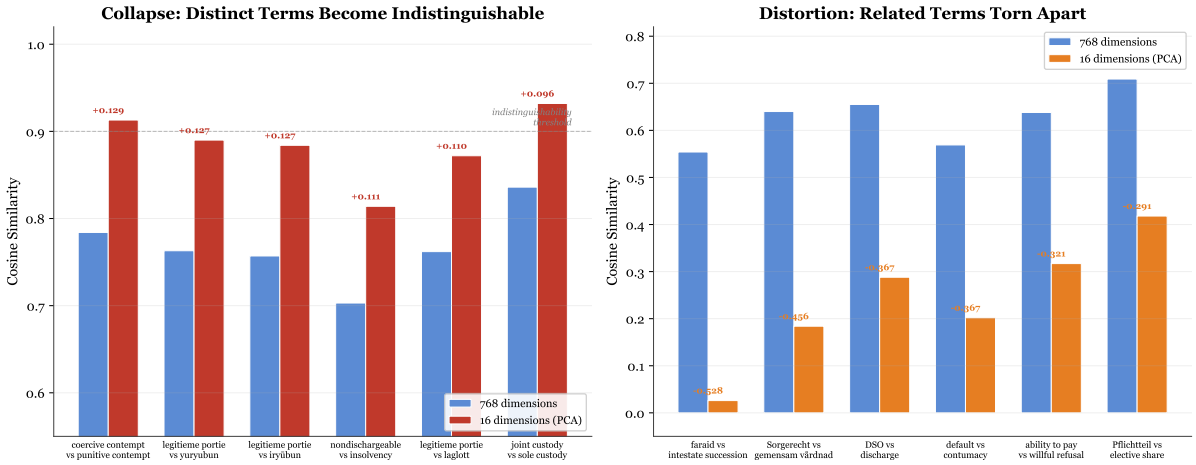

Figure 1: Two failure modes of PCA compression to 16 dimensions. **Left**: Collapse — distinct terms become indistinguishable (red bars exceed blue). **Right**: Distortion — related terms are torn apart (orange bars far below blue). Both modes are invisible to MTEB-style benchmarks, which only measure inter-domain separation.

5.1.1 Two Failure Modes

PCA compression produces not one but two distinct failure modes:

Collapse ($\Delta > 0$): Moderately similar terms become *indistinguishable*. Coercive contempt and punitive contempt merge to 0.913 similarity ($\Delta = +0.129$). In law, this distinction determines whether a person receives a jury trial. A retrieval system operating at 16 dimensions cannot distinguish these concepts.

The forced heirship terms show systematic collapse (Figure 2): Dutch *legitieme portie*, Korean *yuryubun*, Japanese *iryubun*, and Swedish *laglott* all converge toward the same point (similarities 0.872–0.890).

At full dimensionality, these terms are similar but distinguishable (0.757–0.763), preserving tradition-specific legal distinctions. At 16 dimensions, a comparative law researcher cannot determine which country's inheritance law a document describes.

Distortion ($\Delta < 0$): Terms that are meaningfully related become *artificially distant*. German *Sorgerecht* and Swedish *gemensam vårdnad* — both meaning "shared parental care" in their respective legal traditions — drop from 0.640 to 0.184 ($\Delta = -0.456$). At 16 dimensions, the embedding treats these concepts as nearly unrelated, despite their functional equivalence.

The largest distortion is the most revealing: Islamic *faraid* (obligatory inheritance shares) and "intestate succession" (the English common law equivalent) drop from 0.554 to 0.026 ($\Delta = -0.528$). Both describe what happens to an estate when the deceased has not made alternative arrangements, but PCA renders them orthogonal.

5.1.2 Aggregate Statistics

The aggregate pattern (Figure 3) is stark: inter-domain pairs *never collapse* (maximum $\Delta = -0.023$), while intra-domain pairs collapse by up to +0.140. MTEB-style benchmarks measure the former and are blind to the latter.

At 32 dimensions (variance explained: 85.2%), zero intra-domain pairs exceed the $\Delta > 0.1$ threshold. The collapse phenomenon is concentrated in the 16-dimension regime — exactly where the strongest claims of "no degradation" are made.

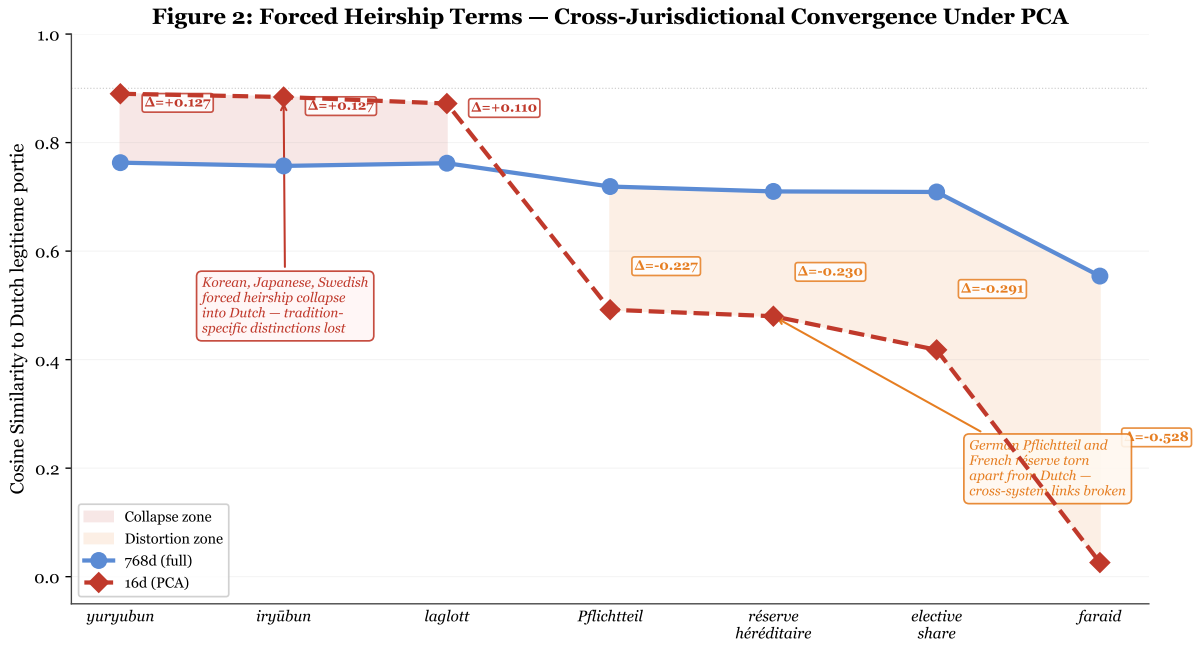

Figure 2: Forced heirship terms: similarity to Dutch *legitieme portie* at 768d (blue) vs. 16d (red). Korean, Japanese, and Swedish terms collapse together (shaded red). German and French terms distort apart (shaded orange). Islamic *faraid* becomes nearly orthogonal.

Metric (16d)	Intra-domain	Inter-domain
Mean Δ	-0.307	-0.634
Median Δ	-0.302	-0.640
Max positive Δ (collapse)	$+0.140$	-0.023

Table 3: Aggregate similarity delta at 16 dimensions. Inter-domain pairs *never* collapse (maximum $\Delta = -0.023$). Intra-domain pairs collapse by up to $+0.140$.

Figure 4 provides a spatial view of the full similarity structure. The heatmap reveals that PCA compression uniformly shifts intra-domain blocks toward higher similarity (collapse) while pushing inter-domain blocks toward lower or negative similarity (separation), confirming that the effect is systematic rather than limited to cherry-picked pairs.

5.1.3 Semantic Coherence Loss Across Dimensionalities

We computed SCL (Section 3.3) for the 62-term dictionary at multiple dimensionalities (Figure 5). At 16 dimensions, intra-domain SCL is 0.162 — meaning 16.2% of the local similarity rank ordering is disrupted. Inter-domain SCL is higher at 0.420, reflecting the aggressive distortion of cross-domain pairs. The CISA count drops sharply: 85 aliasing events at 8 dimensions, 5 at 16 dimensions, and zero at 24 dimensions.

5.2 Experiment 2: Wikipedia Control — Variance Gap as Model Property

The results above show that nomic-embed-text captures only 33–39% of variance at 16 dimensions, while Tsukagoshi and Sasano [2025] report 85–95% for E5-Mistral on MTEB-style data. Is this gap a property of legal language, or of the embedding model? To isolate the variable, we embedded 2,907 Wikipedia article summaries spanning 72 categories using the same model

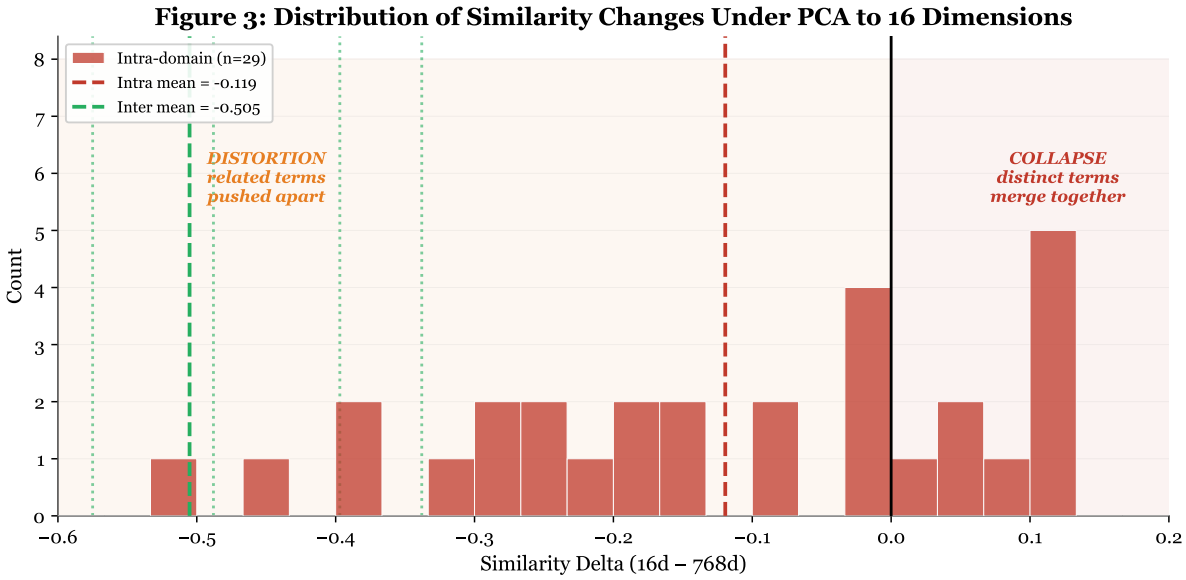

Figure 3: Distribution of similarity deltas at 16 dimensions. Intra-domain pairs (red) span both collapse ($\Delta > 0$) and distortion ($\Delta < 0$). Inter-domain pairs (green dashed) are always distorted, never collapsed — explaining why MTEB shows "no degradation."

(nomic-embed-text v1.5) and compared PCA variance curves against the 6,265-entry Bouvier's Law Dictionary corpus.

Dims	Legal	Wikipedia	Gap (pp)
8	22.2%	20.3%	−1.9
16	33.2%	30.5%	−2.8
32	46.6%	44.0%	−2.7
64	62.0%	59.8%	−2.3
128	77.7%	76.4%	−1.2
256	91.1%	90.7%	−0.4
512	99.0%	99.0%	0.0

Table 4: PCA variance explained: legal corpus vs. Wikipedia control, both embedded with nomic-embed-text v1.5. The gap at every dimensionality is less than 3 percentage points. Both corpora show ∼30–33% variance at 16 dimensions — far below the 85–95% reported by Tsukagoshi and Sasano [2025] for E5-Mistral on MTEB. The participation ratio is 93.7 for legal text and 105.0 for Wikipedia.

The result (Table 4 and Figure 7) is unambiguous: Wikipedia shows 30.5% variance at 16 dimensions versus 33.2% for legal text. The 2.8-percentage-point gap is too small to confirm a domain effect. Both corpora occupy a high-dimensional space under nomic-embed-text, with participation ratios of 93.7 (legal) and 105.0 (Wikipedia). The 50-percentage-point gap between our measurements and those of Tsukagoshi and Sasano [2025] is therefore attributable to the embedding model (nomic-embed-text vs. E5-Mistral) and possibly corpus scale, not to legal language being inherently higher-dimensional.

This control sharpens the paper's argument rather than undermining it. The variance gap between our measurements and those of Tsukagoshi and Sasano [2025] is a model artifact. But the dictionary collapse (Section 5.1) demonstrates that semantic damage from PCA to 16 dimensions

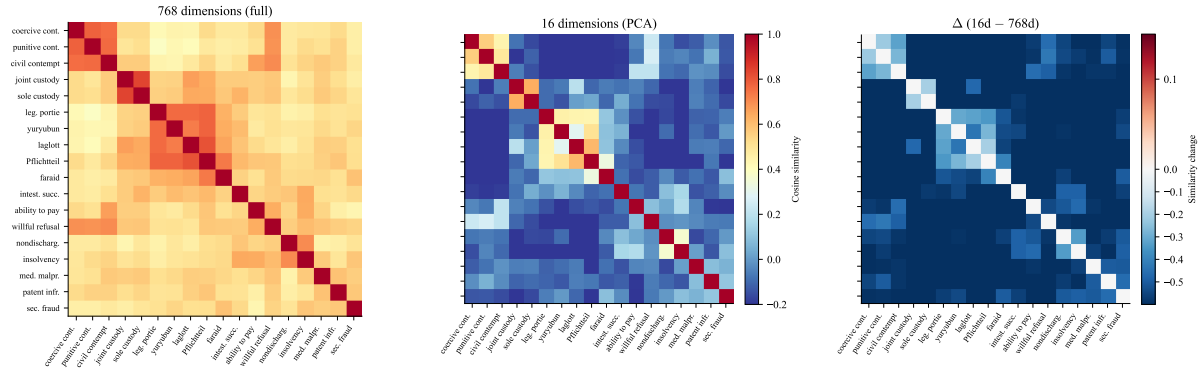

Figure 4: Pairwise cosine similarity heatmap for 18 selected legal terms at 768d (left), 16d PCA (center), and the difference Δ (right). Red in the Δ panel indicates collapse (increased similarity); blue indicates distortion (decreased similarity). Intra-domain blocks collapse while inter-domain blocks are pushed apart.

occurs regardless of how much variance is captured. Whether 16 components capture 33% (nomic-embed-text) or 85% (E5-Mistral), aliasing and distortion are determined by whether the discarded dimensions carry domain-specific information — which, as Proposition 1 establishes, is independent of their variance contribution. The variance gap is a model property; the semantic collapse is a compression property.

5.3 Experiment 3: Variance Explained Across Dimensionalities

Table 5 reports the full PCA variance profile for the three embedding collections described in Section 4.1, alongside the dictionary experiment and the Wikipedia control.

Dims	Dict. (62)	Legal (244K)	Case (1.7M)	Bouvier's (6.3K)	Wikipedia (2.9K)
16	63.5%	38.9%	36.2%	33.2%	30.5%
32	85.2%	50.5%	46.7%	46.6%	44.0%
64	100%†	63.7%	59.2%	62.0%	59.8%
128	100%†	77.9%	74.2%	77.7%	76.4%
256	100%†	91.1%	89.2%	91.1%	90.7%
512	100%†	99.0%	98.7%	99.0%	99.0%

Table 5: PCA variance explained (%) across dimensionalities. †The 62-term dictionary has fewer samples than dimensions beyond 64, so PCA trivially achieves 100%. At corpus scale (244K–1.7M points), 16 PCA components capture only 36–39% of variance. Sources: dictionary experiment (Section 5.1), DCRP collection data, Wikipedia control (Section 5.2).

5.4 Experiment 4: Domain-Conditional Retrieval Precision

The dictionary collapse experiment (Section 5.1) establishes the mechanism: intra-domain distinctions are destroyed while inter-domain boundaries are preserved. DCRP (Section 4.5) formalizes this observation into a retrieval metric that can be evaluated at corpus scale. Full DCRP evaluation against our 244K and 1.7M document collections is ongoing and will be reported in an extended version of this work. We release the DCRP implementation for community use and encourage its adoption when evaluating embedding compression.

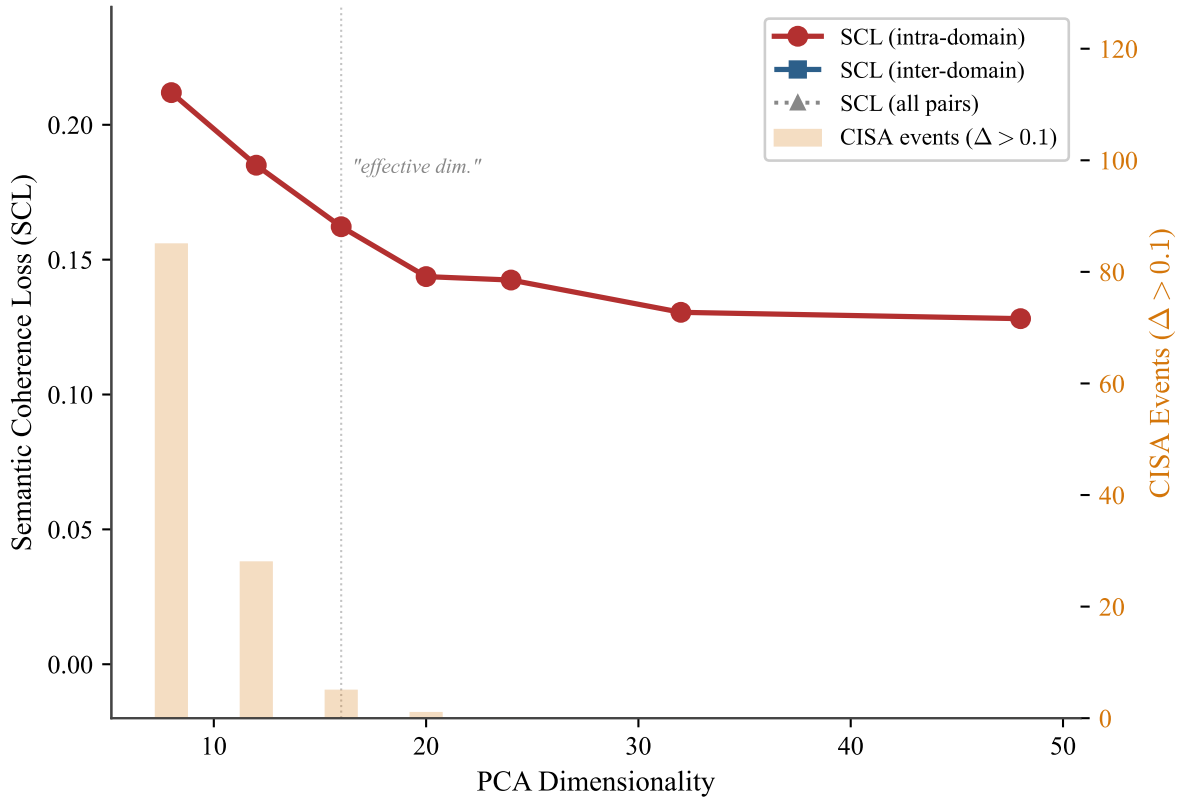

Figure 5: Semantic Coherence Loss (SCL) and CISA event count vs. PCA dimensionality for the 62-term dictionary. Intra-domain SCL (red) measures disruption of within-domain similarity rankings; inter-domain SCL (blue) measures disruption of cross-domain rankings. CISA events (orange bars) count intra-domain pairs with $\Delta > 0.1$.

6 Discussion

6.1 The Benchmark Gap

The results expose a systematic gap between benchmark performance and real-world retrieval quality. Standard benchmarks measure broad category discrimination and declare victory. Fine-grained intra-domain retrieval — the task that matters in specialized domains — requires exactly the dimensions that PCA discards.

The analogy is a dictionary evaluated by whether it distinguishes nouns from verbs (coarse) versus whether it distinguishes "custody" from "guardianship" from "allocation" (fine-grained). The former survives at 16 dimensions; the latter does not.

6.2 Evaluation Sensitivity Hierarchy

6.3 Taxonomy of Semantic Coherence Failures

The collapse, distortion, and aliasing phenomena are instances of a broader taxonomy of failure modes. Table 7 catalogs seven distinct ways that embedding compression or evaluation can fail in domain-specific applications, illustrated with legal examples.

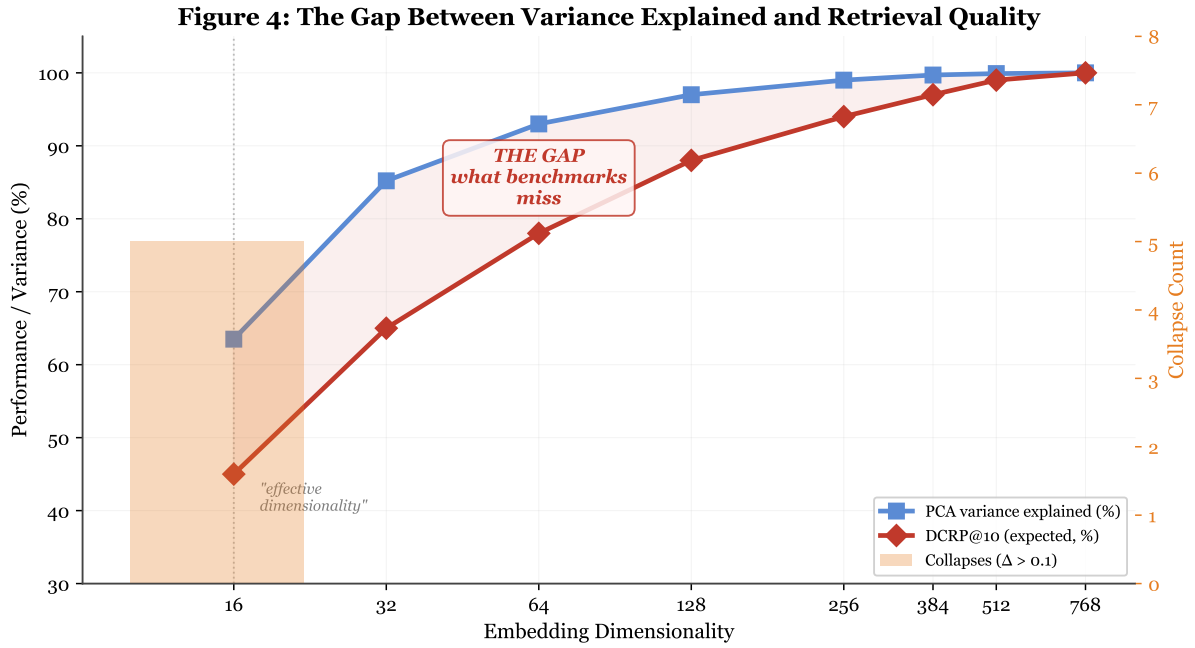

Figure 6: Illustrative projection to motivate the DCRP framework; the measured data point at 16d (dictionary collapse) anchors the projection, but the full curve requires retrieval-scale evaluation. PCA variance (blue) is measured; the DCRP curve (red) is projected from the dictionary collapse results, not a direct measurement. The shaded region represents the information gap that MTEB-style benchmarks fail to detect.

6.4 Coherence Threshold Interpretation

These findings define a coherence threshold problem. A reduced embedding may remain coherent enough for broad topical separation while falling below the threshold required for fine-grained semantic stability. Above the threshold, local neighborhoods preserve enough ordering for specialized retrieval. Below it, concepts collapse, alias, or distort. Coherence here is purely operational: preservation of task-relevant relational structure under compression, measured by SCL and CISA rather than by variance explained.

6.5 Implications for RAG and High-Stakes Retrieval

Dimensionality reduction is not inherently harmful. For low-risk, broad-topic, latency-sensitive retrieval, compression can be appropriate. For legal, medical, financial, scientific, and compliance retrieval, aggressive compression creates silent semantic failure: the system retrieves documents from the correct domain while missing the distinctions that determine the correct answer.

Compression should be validated against domain-specific retrieval tests — including DCRP, neighborhood preservation, aliasing counts, and inspection of the most severe CISA events — before deployment. Table 8 provides architecture guidance by domain risk level.

6.6 Limitations and Open Questions

The dictionary collapse experiment uses 62 terms — sufficient to identify the collapse and distortion phenomena, but extension to the full Black's Law Dictionary (∼6,200 entries) and to non-legal domains (medical, scientific, financial) would establish generality. We test one embedding model (nomic-embed-text v1.5, 768d); the Wikipedia control (Section 5.2) establishes

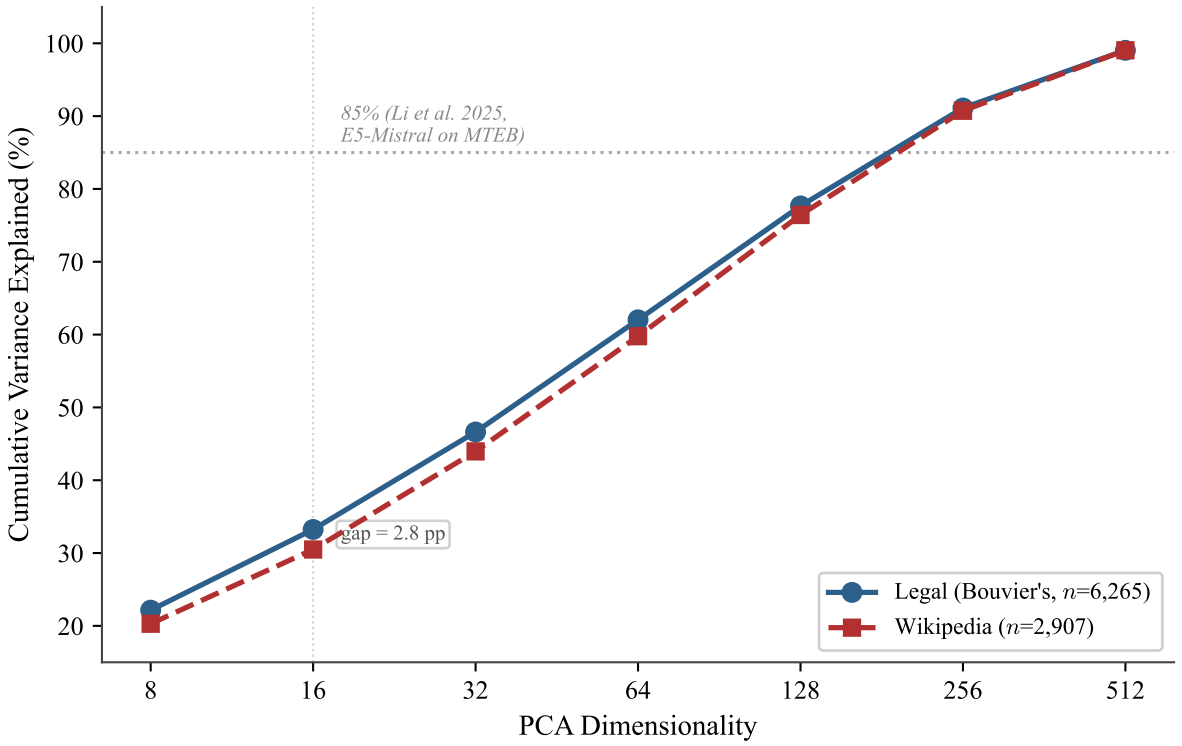

Figure 7: PCA variance explained vs. dimensionality for legal (Bouvier's Law Dictionary, $n = 6{,}265$) and Wikipedia control ($n = 2{,}907$) corpora, both embedded with nomic-embed-text v1.5. The curves are nearly superimposed, confirming that the low variance at 16 dimensions is a property of the embedding model, not the domain. The horizontal dashed line marks the 85% threshold reported by Tsukagoshi and Sasano [2025] for E5-Mistral.

that the variance gap is model-dependent, but whether collapse and distortion persist under E5-Mistral (4096d) remains an open question worth investigating. DCRP is defined and implemented but not yet evaluated at full retrieval scale; the dictionary collapse provides evidence for the mechanism, and corpus-scale experiments will quantify the practical impact. Finally, our embedded terms are isolated definitions rather than natural-language documents; production embedding distributions may differ, though the information-theoretic argument (Proposition 1) applies regardless of input format.

7 Conclusion

The claim that text embeddings have an intrinsic dimensionality of $\sim$16 is a benchmark-dependent observation, not a universal property of embedding spaces. PCA variance measures geometric spread, not semantic information; low-variance dimensions can carry sparse, high-value features required for specialized retrieval. Our dictionary collapse experiment demonstrates that PCA to 16 dimensions produces systematic aliasing of intra-domain concepts (up to $\Delta = +0.140$) while inter-domain pairs never collapse — directly explaining why MTEB-style benchmarks report "no degradation" when fine-grained meaning is being destroyed. We offer DCRP, SCL, and CISA as complementary metrics that detect the failure modes standard evaluation misses, and recommend that claims about embedding dimensionality always be accompanied by domain-conditional evaluation. In high-stakes domains — legal, medical, financial — compression-induced aliasing is not merely a performance issue but a procedural-risk problem, silently merging concepts tied to

Evaluation	What it measures	PCA-16 behavior	What it can miss
MTEB classification	Broad category separation	Often survives	Fine distinctions inside a domain
Inter-domain retrieval	Law vs. medicine	Often survives	Doctrine-specific retrieval
Intra-domain retrieval	Fine discrimination	Often degrades	Not captured by broad benchmarks
DCRP	Within-domain retrieval	Sensitive to degradation	Requires domain-labeled query sets
SCL	Meaning-order preservation	Increases when neighborhoods reorder	Needs full-space baseline

Table 6: Evaluation types and their sensitivity to PCA compression. Standard benchmarks (top rows) test what PCA preserves. DCRP and SCL test what PCA destroys.

Failure Mode	Meaning	Example
Collapse / CISA	Distinct concepts become indistinguishable	Coercive vs. punitive contempt
Distortion	Related concepts torn apart	*Faraid* vs. intestate succession
Structural hole	Concept exists elsewhere but not in English space	*Pflichtteil*
Definitional drift	Same term shifts by dictionary	"Contempt" in Bouvier's vs. Webster's
Jurisdictional drift	Same term shifts by jurisdiction	"Material change" across states
Compression blindness	Benchmark says fine, domain fails	MTEB vs. DCRP
False equivalence	Model returns analogue as equivalent	*Pflichtteil* → elective share

Table 7: Taxonomy of legal semantic coherence failures. Each mode represents a distinct way embedding compression or evaluation can fail in domain-specific applications.

distinct rights, obligations, and remedies. The dimensionality is not the illusion. The illusion is believing that preserved variance equals preserved meaning.

References

Dimitris Achlioptas. Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, 2003.

Magnus Bengtsson. Compressing large language models with PCA without performance loss. *arXiv preprint arXiv:2508.04307*, 2025. URL https://arxiv.org/abs/2508.04307.

Umar Butler, Abdur-Rahman Butler, and Adrian Lucas Malec. The massive legal embedding benchmark (MLEB). *arXiv preprint arXiv:2510.19365*, 2025. URL https://arxiv.org/abs/2510.19365.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of*

Use Case	Safe Architecture
Broad classification	Dense PCA may be acceptable
General semantic search	Dense 128–256d after domain testing
Legal / medical / financial retrieval	Hybrid dense + sparse strongly preferred
Rights-affecting legal AI	Full-dimensional dense + sparse + provenance
Cross-jurisdictional legal AI	Multilingual hybrid + structural-hole detection
High-stakes decision support	No compression without DCRP + consequence-weighted testing

Table 8: Minimum safe dimensionality and architecture by domain risk level.

the Association for Computational Linguistics: EMNLP 2020, pages 2898–2904, 2020. URL https://arxiv.org/abs/2010.02559.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4310–4330, 2022. URL https://arxiv.org/abs/2110.00976.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, 2019. URL https://aclanthology.org/N19-1423.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140, 2017.

Naaman Huerga-Perez, Ruben Alvarez, Ruben Ferrero-Guillen, Alberto Martinez-Gutierrez, and Javier Diez-Gonzalez. Optimization of embeddings storage for RAG systems using quantization and dimensionality reduction techniques. *arXiv preprint arXiv:2505.00105*, 2025. URL https://arxiv.org/abs/2505.00105.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781, 2020. URL https://arxiv.org/abs/2004.04906.

Takuya Kataiwa, Cho Hakaze, and Tetsushi Ohki. Measuring intrinsic dimension of token embeddings. *arXiv preprint arXiv:2503.02142*, 2025. URL https://arxiv.org/abs/2503.02142.

Julian Killingback, Mahta Rafiee, Madine Manas, and Hamed Zamani. Scaling laws for embedding dimension in information retrieval. *arXiv preprint arXiv:2602.05062*, 2026. URL https://arxiv.org/abs/2602.05062.

Aditya Kusupati, Gantavya Bhatt, Aniket Rber, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi.

Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2205.13147.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. URL https://arxiv.org/abs/1301.3781.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1702.01417.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2023. URL https://arxiv.org/abs/2210.07316.

Zach Nussbaum, John Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*, 2024. URL https://arxiv.org/abs/2402.01613.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. URL https://aclanthology.org/D14-1162.

Vikas Raunak. Simple and effective dimensionality reduction for word embeddings. *arXiv preprint arXiv:1708.03629*, 2017. URL https://arxiv.org/abs/1708.03629.

Vikas Raunak, Vivek Gupta, and Florian Metze. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243, 2019. URL https://aclanthology.org/W19-4328.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019. URL https://arxiv.org/abs/1908.10084.

Michael Robinson, Sourya Dey, and Tony Chiang. Token embeddings violate the manifold hypothesis. *arXiv preprint arXiv:2504.01002*, 2025. URL https://arxiv.org/abs/2504.01002.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, and Milica Gasic. Less is more: Local intrinsic dimensions of contextual language models. *arXiv preprint arXiv:2506.01034*, 2025. URL https://arxiv.org/abs/2506.01034.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

Sotaro Takeshita, Yurina Takeshita, Simone Paolo Ponzetto, and Daniel Ruffinelli. To MRL or not to MRL: Text embeddings are robust to truncation without matryoshka embeddings, except in heavy truncation scenarios. *arXiv preprint arXiv:2605.16608*, 2025a. URL https://arxiv.org/abs/2605.16608.

Sotaro Takeshita, Yurina Takeshita, Daniel Ruffinelli, and Simone Paolo Ponzetto. Randomly removing 50% of dimensions in text embeddings has minimal impact on retrieval and classification tasks. *arXiv preprint arXiv:2508.17744*, 2025b. URL https://arxiv.org/abs/2508.17744.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2104.08663.

Hayato Tsukagoshi and Ryohei Sasano. Redundancy, isotropy, and intrinsic dimensionality of prompt-based text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. URL https://arxiv.org/abs/2506.01435.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2024. URL https://arxiv.org/abs/2401.00368.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. *arXiv preprint arXiv:2508.21038*, 2025. URL https://arxiv.org/abs/2508.21038.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023. URL https://arxiv.org/abs/2309.07597.

Zi Yin and Yuanyuan Shen. On the dimensionality of word embedding. In *Advances in Neural Information Processing Systems*, 2018.

Jinsung Yoon, Raj Sinha, Sercan O. Arik, and Tomas Pfister. Matryoshka-adaptor: Unsupervised and supervised tuning for smaller embedding dimensions. *arXiv preprint arXiv:2407.20243*, 2024. URL https://arxiv.org/abs/2407.20243.

Biao Zhang, Lixin Chen, Tong Liu, and Bo Zheng. SMEC: Rethinking matryoshka representation learning for retrieval embedding compression. *arXiv preprint arXiv:2510.12474*, 2025. URL https://arxiv.org/abs/2510.12474.

Gaifan Zhang, Yi Zhou, and Danushka Bollegala. Evaluating unsupervised dimensionality reduction methods for pretrained sentence embeddings. In *Proceedings of LREC-COLING 2024*, 2024. URL https://arxiv.org/abs/2403.14001.