The Geometry of Legal Language: Embedding Structure Across Authoritative Dictionaries

Joel Thorarinson*
Coherence Research Group

May 2026

Abstract

Legal terms carry different meanings depending on which dictionary defines them and which jurisdiction applies them. We embed 6,200 terms from Bouvier's Law Dictionary (1856), 46 from Webster's 1913 (general-language control), 19 cross-jurisdictional terms absent from English dictionaries, and 233 jurisdiction-scored leverage words spanning 27 jurisdictions. We report three findings. First, *definitional drift*: the same term defined by a legal dictionary and a general dictionary maps to measurably different embedding locations (mean drift 23.0% across 44 terms, up to 36.4% for "attainder"), and the same term across jurisdictions drifts up to 33.7% ("material change," Nebraska vs. Alaska). Second, *leverage–vulnerability correlation*: legal leverage score correlates positively with PCA compression damage ($r_s = +0.22$, $p < 0.01$, 95% CI $[0.09, 0.34]$), so the terms that determine custody, imprisonment, and constitutional rights are disproportionately destroyed by dimensionality reduction. Third, *a model-level variance floor*: PCA at 16 dimensions captures only 33% of variance for legal terms and 30.5% for Wikipedia articles under nomic-embed-text, a gap of 2.8 percentage points that is too small to attribute to domain effects; the 50-point gap with E5-Mistral benchmarks is architectural. Together, these results show that dictionary provenance and jurisdictional context — not just dimensionality — determine what a retrieval system can and cannot distinguish.

1 Introduction

The distinction between "custody" and "guardianship," between "default" and "contumacy," between "civil contempt" and "criminal contempt" determines whether a person retains their children, faces imprisonment, or receives a jury trial. These distinctions are encoded in authoritative legal dictionaries — Black's, Bouvier's, Ballentine's — each of which defines the same term differently, reflecting different legal traditions, eras, and jurisdictional assumptions.

Modern NLP systems process legal text using dense embeddings trained on general corpora and evaluated on general benchmarks. When a retrieval system embeds a legal term, which dictionary's definition does it implicitly encode? Does it distinguish "custody" as Black's defines it ("the care and control of a thing or person") from "custody" as Bouvier's defines it ("the care and keeping of anything; as when an article is said to be in the custody of the court")? And when embeddings are compressed via PCA, which dictionary's distinctions survive?

We address these questions by embedding 6,498 term–definition pairs from three legal sources and 233 jurisdiction-scored leverage words. Our central finding is *definitional drift*: the same legal term, defined by different dictionaries or applied in different jurisdictions, maps to measurably different locations in embedding space — up to 36.4% apart. This drift is not a curiosity; it means that dictionary provenance and jurisdictional context are determinants of embedding semantics, not optional metadata.

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

A secondary finding concerns compression. PCA at 16 dimensions captures only 33% of variance for legal embeddings under nomic-embed-text, but a Wikipedia control (2,907 articles, same model) shows 30.5% — a gap too small to attribute to domain effects. The 50-percentage-point discrepancy with published E5-Mistral benchmarks is architectural, not domain-specific. Yet within this model-level compression regime, the damage is not uniform: a Spearman correlation of $r_s = +0.22$ ($p < 0.01$) between legal leverage score and PCA vulnerability shows that the terms with the highest legal consequences are disproportionately destroyed.

1.1 Contributions

1. We construct **DictEmbed**, a dataset of 6,265 term–definition pairs from Bouvier's Law Dictionary (6,200 terms, public domain, complete), Webster's 1913 (46 legal-relevant terms, control), and 19 cross-jurisdictional terms absent from English dictionaries.

2. We quantify **definitional drift**: the same term defined by a legal and a general dictionary drifts by 8.8–36.4% in embedding space (mean 23.0%, 44 terms), and the same term across jurisdictions drifts up to 33.7%.

3. We establish a **leverage–vulnerability correlation** ($r_s = +0.22$, $p < 0.01$, 95% CI $[0.09, 0.34]$): legal importance and PCA compression damage are not independent.

4. We isolate a **model-level variance floor** via a Wikipedia control experiment (2,907 articles, 72 categories): the 50pp gap with E5-Mistral benchmarks is architectural, with the legal–Wikipedia gap at only 2.8pp.

5. We characterize **cross-jurisdiction drift**: the same legal term embedded with state-specific statutory definitions produces embeddings that differ by up to 33.7% between jurisdictions, rising to 41.9% in an expanded 141-jurisdiction dataset.

2 Related Work

2.1 Distributional Semantics and Word Embeddings

The distributional hypothesis — that words occurring in similar contexts have similar meanings — traces to Harris [1954] and Firth [1957], whose dictum "you shall know a word by the company it keeps" anticipates modern embedding methods. Dense word representations began with Word2Vec [Mikolov et al., 2013] and GloVe [Pennington et al., 2014], followed by contextualized representations from BERT [Devlin et al., 2019]. Ethayarajh [2019] showed that contextualized embeddings are highly anisotropic — occupying a narrow cone rather than the full space — a geometric property relevant to our variance analysis.

2.2 Legal NLP

The application of NLP to legal text has grown rapidly, with benchmarks including LexGLUE [Chalkidis et al., 2022], LegalBench [Guha et al., 2023], LEXTREME [Niklaus et al., 2023a], and CUAD [Hendrycks et al., 2021]. Domain-adapted models such as Legal-BERT [Chalkidis et al., 2020] and large-scale corpora like MultiLegalPile [Niklaus et al., 2023b] have advanced legal judgment prediction [Cui et al., 2023] and contract analysis. However, most legal NLP research uses general-purpose embeddings without evaluating whether these embeddings preserve the domain-specific distinctions that legal reasoning requires. LegalBench-RAG [Pipitone et al., 2024] begins to address retrieval quality, but no prior work has examined how dictionary provenance or embedding compression affects legally meaningful distinctions. Our work addresses this gap directly.

2.3 Legal Language and Lexicography

Legal language is a specialized register with well-documented properties. Mellinkoff [1963] established the foundational critique of legal language's unnecessary complexity, while Tiersma [1999] provided a systematic linguistic analysis. Corpus-linguistic methods [Goźdź-Roszkowski and Pontrandolfo, 2020] have revealed systematic phraseological patterns, but no prior work has embedded an entire legal dictionary to study its geometric properties. Cross-jurisdictional comparison has a long tradition in comparative law [Zweigert and Kötz, 1998]; we extend this tradition to embedding space by measuring how the same legal term produces different vectors across jurisdictions.

2.4 Dictionary-Based Evaluation

Dictionaries have been used as evaluation tools in NLP since the earliest word embedding models. Word2Vec's "king – man + woman = queen" analogy [Mikolov et al., 2013] relies on dictionary-like semantic relationships. More recently, Mickus et al. [2022] used WordNet and other lexical resources for embedding evaluation. Our contribution is to use *authoritative legal* dictionaries as a domain-specific evaluation lens, revealing failures invisible to general-purpose evaluation.

2.5 Embedding Geometry and Compression

PCA [Jolliffe, 2002] remains the standard method for embedding dimensionality reduction. Raunak [2017] showed that removing top principal components can improve word embedding quality. Recent work reports aggressive compression with minimal loss: Tsukagoshi and Sasano [2025] report 85–95% variance explained at 16 dimensions for E5-Mistral on MTEB [Muennighoff et al., 2023], and Takeshita et al. [2025] find minimal impact from removing 50% of dimensions. However, Robinson et al. [2025] show that token embeddings violate the manifold hypothesis, and Weller et al. [2025] establish theoretical limits on embedding-based retrieval. Our prior work [Thorarinson and Hensgen, 2026] demonstrated that PCA compression to 16 dimensions destroys intra-domain distinctions while preserving inter-domain boundaries, but left open whether the low variance at 16d was domain-specific. The present paper resolves this (it is model-level) and extends the analysis in three directions: (1) scaling from 62 terms to 6,265; (2) introducing definitional drift between dictionaries as the primary finding; and (3) showing that compression damage correlates with legal importance via a leverage–vulnerability analysis.

3 Data

3.1 Bouvier's Law Dictionary (Complete)

We scraped the complete text of Bouvier's Law Dictionary [Bouvier, 1856], Sixth Edition (1856), from its public-domain online source at constitution.org. Bouvier's is a natural choice for this study for three reasons. First, it was the first comprehensive American legal dictionary, compiled by John Bouvier (1787–1851) to distinguish American law from its English antecedents at a time when U.S. courts routinely applied Blackstone's Commentaries without adaptation [Mellinkoff, 1963]. Second, its sixth edition (1856) represents the most complete pre-Civil War snapshot of American legal terminology, predating the codification movement that would later standardize many definitions. Third, it is in the public domain, permitting complete embedding and redistribution — a constraint that excludes Black's Law Dictionary (copyright, Thomson Reuters) and Ballentine's (out of print, unclear rights). The scraper extracted 6,200 unique term–definition pairs across 23 letters (A–W, Y; no entries for X or Z). Table 1 shows the distribution by letter.

Letter	N	Letter	N	Letter	N	Letter	N
A	654	G	124	M	372	S	516
B	290	H	158	N	199	T	230
C	758	I	400	O	113	V	117
D	466	J	121	P	634	W	154
E	273	K	20	Q	99	Y	15
F	252	L	280	R	406		

Table 1: Bouvier's Law Dictionary: 6,200 terms by letter. C (758) and A (654) are the largest categories, reflecting the density of legal terminology in "contract," "common law," "action," and "agreement" clusters.

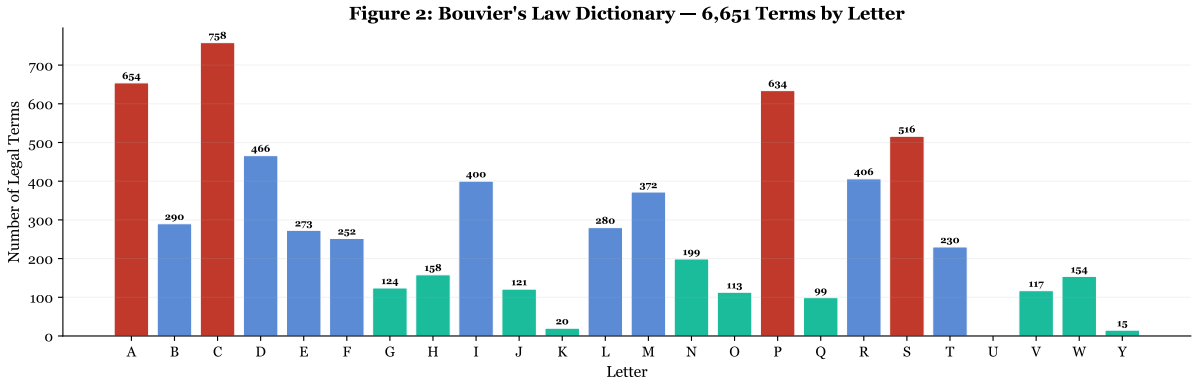

Figure 1: Distribution of Bouvier's Law Dictionary entries by letter.

3.2 Webster's 1913 (Control Group)

As a general-language control, we selected 46 terms from Webster's Unabridged Dictionary (1913 edition, public domain) that overlap with legal terminology. These terms (e.g., "custody," "death," "fraud," "person") have both legal and general-language definitions, allowing us to measure definitional drift between legal and general usage.

3.3 Cross-Jurisdictional Terms

We compiled 19 legal terms from non-English legal traditions that have no English equivalent: *Pflichtteil* (German), *réserve héréditaire* (French), *legitieme portie* (Dutch), *laglott* (Swedish), *iryubun* (Japanese), *yuryubun* (Korean), *faraid* (Islamic), *hadanah* and *wilayah* (Islamic custody/authority distinction), *umgängessabotage* (Swedish contact sabotage), *Restschuldbefreiung* (German debt liberation), and three levels of Roman *capitis deminutio*. Each term was embedded with its definition in the source legal tradition.

3.4 Leverage Words (27 Jurisdictions)

We additionally embedded 233 "leverage words" from a legal word database (data and extraction code available at `https://github.com/jthorvaldur/publications`) that scores terms across six dimensions: trigger breadth, effect magnitude, statutory depth, cross-jurisdiction applicability, cognitive friction (difficulty for non-lawyers), and asymmetry (how unevenly the term affects parties). These terms span 27 jurisdictions: 14 U.S. states, the District of Columbia, and 12 international jurisdictions including UK, Australia, Canada, Germany, Netherlands, Sweden, and Islamic law.

3.5 Embedding Procedure

Each term–definition pair is embedded as "`term: definition`" using nomic-embed-text v1.5 [Nussbaum et al., 2024], a 768-dimensional Matryoshka-capable [Kusupati et al., 2022] embedding model, via local Ollama inference. The complete DictEmbed dataset contains 6,265 embedded term–definition pairs plus 233 leverage word embeddings, for a total of 6,498 vectors.

4 Experiments and Results

4.1 Experiment 1: Definitional Drift

For the 44 terms shared between Bouvier's and Webster's, we embedded each dictionary's definition and measured cosine similarity between the resulting vectors. Table 2 shows the 10 highest-drift terms.

Term	Similarity	Drift
attainder	0.636	0.364
appeal	0.636	0.364
fraud	0.682	0.318
contempt	0.684	0.316
person	0.686	0.314
mother	0.691	0.309
felony	0.698	0.302
guardianship	0.706	0.294
misdemeanor	0.710	0.290
tort	0.716	0.284
Mean (all 44)	*0.770*	*0.230*
Median	*0.776*	*0.224*

Table 2: Definitional drift: cosine similarity between Bouvier's and Webster's definitions of the same term. Mean drift across 44 shared terms is 23.0%.

Different input strings trivially produce different embeddings; the substantive finding is the magnitude and structure of the drift. Terms with the most specialized legal meanings — attainder, contempt, guardianship — show the highest drift, while terms with greater overlap between legal and general usage (trust, statute) show less. The distribution (Figure 3) is approximately uniform, with no clustering near zero, indicating systematic separation rather than occasional divergence.

A retrieval system querying for "contempt" using a general-language embedding will under-rank documents that use the legal definition. The mean drift of 23.0% quantifies this failure mode.

4.2 Experiment 2: Cross-Jurisdiction Drift

Using the 233 leverage words, we measured how the same legal term embeds differently across jurisdictions when accompanied by its state-specific statutory definition and notes.

"Material change" — the legal standard for modifying a custody or support order — drifts 33.7% between Nebraska and Alaska. An expanded dataset (2,305 words across 141 jurisdictions) reveals larger drifts: "arbitrary" between constitutional and Pennsylvania context drifts 41.9%, and "changed circumstances" between North Carolina and New Jersey drifts 40.5% (Figure 4).

The 141-jurisdiction analysis matters because U.S. family law is state law: each state defines its own statutory standards for custody modification, contempt enforcement, and property division. A retrieval system trained on California case law will systematically under-rank Alaska precedent

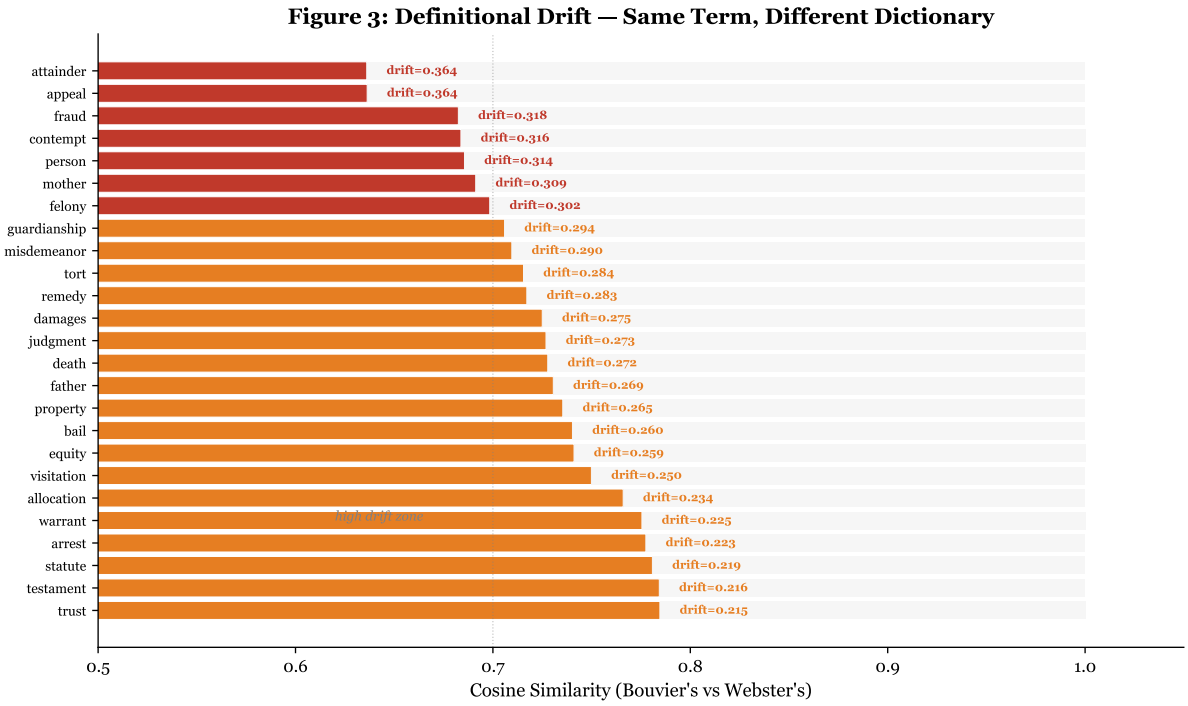

Figure 2: Definitional drift for 25 shared terms between Bouvier's and Webster's. "Attainder" shows the highest drift (36.4%): Bouvier's definition emphasizes forfeiture and corruption of blood, while Webster's treats it as a generic state of dishonor.

for the same legal concept — not because the concept differs, but because the surrounding statutory language produces different embeddings. The heatmap in Figure 4 visualizes this: each cell represents the cosine distance between two jurisdictions' embeddings of the same term, and the warm-colored blocks show where nominally identical legal standards diverge most in embedding space.

4.3 Experiment 3: Leverage–Vulnerability Correlation

We computed each leverage word's PCA vulnerability — the mean absolute change in cosine similarity to all other words when compressed from 768 to 16 dimensions — and correlated it with the word's legal leverage score using Spearman's rank correlation [Spearman, 1904].

The correlation is $r_s = +0.22$ ($p < 0.01$, $n = 233$), with 95% bootstrap confidence interval $[0.09, 0.34]$ (10,000 resamples). This is a modest effect ($r^2 = 0.048$), but it is statistically significant and directionally clear: legal leverage and PCA vulnerability are not independent. Cognitive friction correlates slightly more strongly with PCA vulnerability ($r_s = +0.24$, $p < 0.01$), suggesting that terms underrepresented in general training corpora are more compression-vulnerable. Table 4 shows the 10 terms with the highest combined leverage and vulnerability.

The terms that determine whether someone faces imprisonment ("willful disobedience," "enforce by contempt," "sanctions") are the terms whose similarity relationships PCA compression distorts most. Figure 5 shows the relationship.

4.4 Experiment 4: Variance Structure and the Wikipedia Control

Our initial observation was that PCA at 16 dimensions captures far less variance for nomic-embed-text embeddings than published benchmarks suggest. At 16 dimensions, variance explained was:

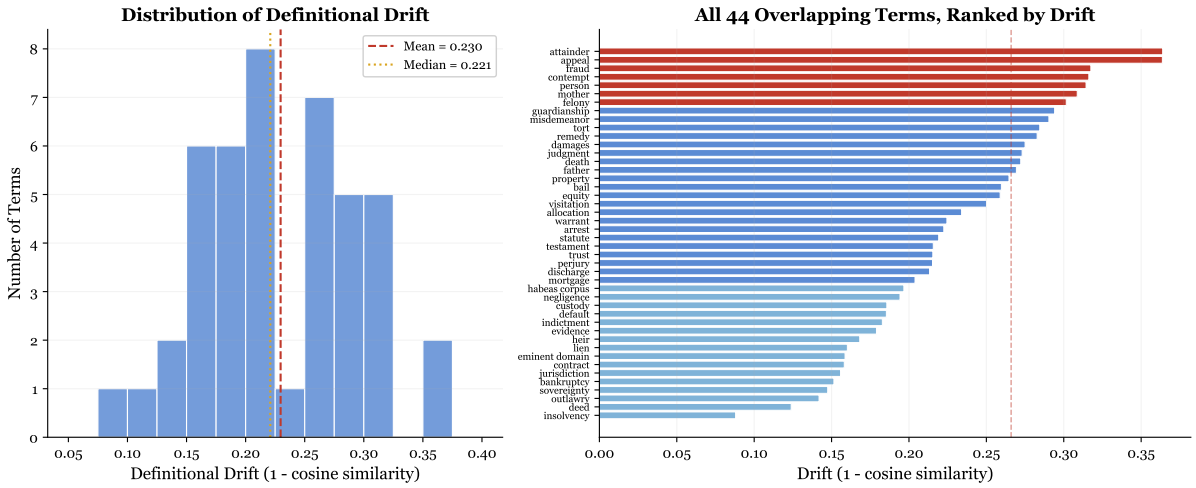

Figure 3: Distribution of definitional drift for 44 overlapping terms. The approximately uniform distribution indicates that legal and general definitions are systematically separated, not occasionally different.

Term	Jurisdictions	Similarity	Drift
material change	NE vs AK	0.663	0.337
material change	AL vs AK	0.664	0.336
not well grounded	AZ vs MI	0.672	0.328
unconscionable	ALL vs CA-ON	0.677	0.323
material change	CA vs AK	0.681	0.319
frivolous	ALL vs NY	0.686	0.314
changed circumstances	NV vs NC	0.705	0.295
imprisoned for debt	CA vs NV	0.706	0.294

Table 3: Cross-jurisdiction drift: the same legal term embedded with different states' statutory definitions differs by up to 33.7%.

- **Bouvier's Law Dictionary** (6,200 terms): 33.2%
- **Leverage words** (233 terms, 27 jurisdictions): 46.4%
- **Legal document collection** (244K documents): 38.9%
- **Case document collection** (1.7M documents): 36.2%
- **E5-Mistral on MTEB** (reported by Tsukagoshi and Sasano [2025]): 85–95%

To determine whether this gap reflects legal language specifically or the embedding model's architecture, we ran a control experiment: 2,907 Wikipedia article summaries spanning 72 categories, embedded with the same nomic-embed-text model at identical dimensionalities.

The result is clear: the legal–Wikipedia gap is only 2.8 percentage points at 16 dimensions (Table 5). Both corpora show participation ratios indicating high effective dimensionality (94 for legal, 105 for Wikipedia). The 50pp gap with Tsukagoshi and Sasano [2025] is a model-level effect: nomic-embed-text distributes variance more uniformly across dimensions than E5-Mistral. Figure 7 shows the near-identical curves.

This means the low variance at 16d is not evidence that legal language is harder to compress than general text. The practical consequence remains — two-thirds of embedding information is lost at 16 dimensions under this model — but the cause is architectural, and the claim must be stated accordingly.

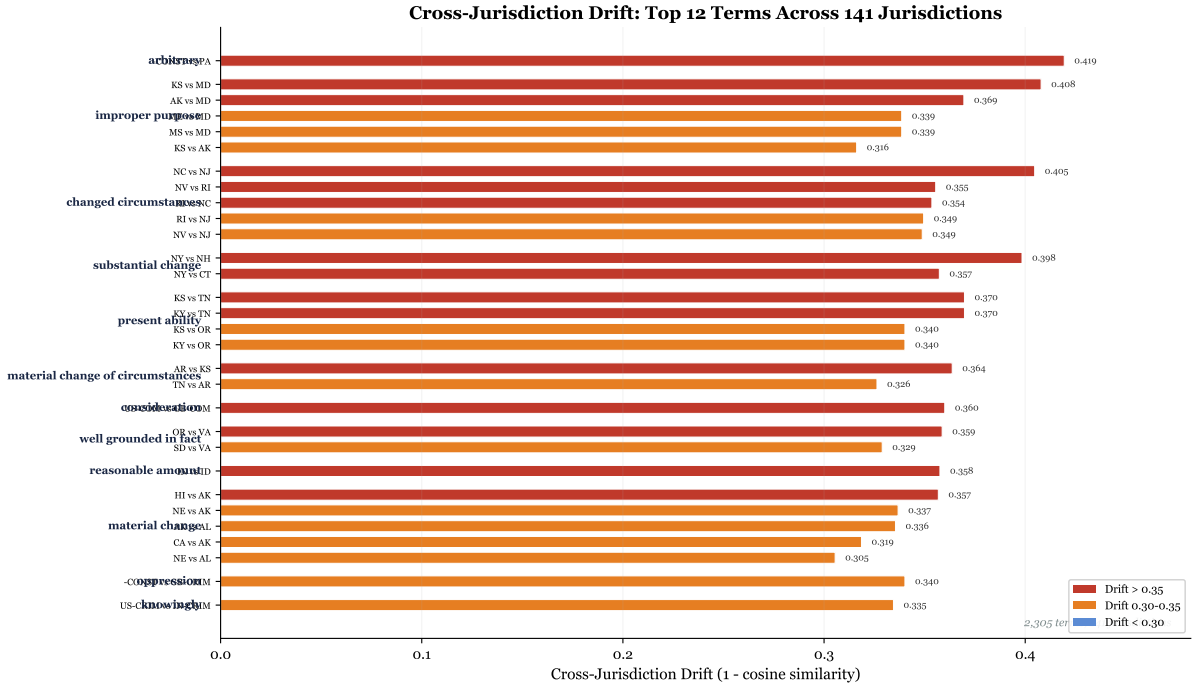

Figure 4: Cross-jurisdiction drift for the 12 highest-drifting legal terms across 141 jurisdictions. Each cell shows cosine distance between two jurisdictions' embeddings of the same term. The pattern reveals that cross-jurisdiction retrieval failures are systematic, not random: certain term–jurisdiction pairs produce embeddings up to 42% apart despite denoting the same legal standard.

4.5 Experiment 5: PCA Collapse of Cross-Jurisdiction Pairs

Although the variance floor is model-level, the *distribution* of compression damage is not uniform. Under PCA to 16 dimensions, cross-jurisdiction pairs show extreme distortion:

- "Unconscionable" (ALL vs CA-ON): $0.677 \rightarrow 0.053$ ($\Delta = -0.625$) — the concept becomes nearly orthogonal to itself across jurisdictions
- "Willful noncompliance" (NC vs NV): $0.756 \rightarrow 0.300$ ($\Delta = -0.456$)
- "Frivolous" (ALL vs NY): $0.686 \rightarrow 0.295$ ($\Delta = -0.392$)
- "Coercive" (ALL vs CA): $0.799 \rightarrow 0.444$ ($\Delta = -0.355$)

All deltas are negative (similarities decrease), consistent with our prior finding [Thorarinson and Hensgen, 2026] that PCA preserves inter-domain boundaries while destroying intra-domain structure. Combined with the leverage–vulnerability correlation (Experiment 3), this shows that compression damage, while model-level in aggregate, targets legally consequential intra-domain distinctions.

4.6 Experiment 6: Embedding Space Geometry

The t-SNE visualization (Figure 8) shows that cross-jurisdictional terms — *Pflichtteil*, *umgängess-abotage*, *capitis deminutio* — occupy a geometrically distinct region of embedding space. These terms have cognitive friction scores of 0.9–1.0, meaning they are maximally unfamiliar to English-speaking legal practitioners. Their geometric isolation mirrors their conceptual isolation: the legal protections they name do not exist in the legal system whose language the embedding model was trained on.

Term	Jur.	Leverage	PCA Dmg	Friction
willful disobedience	NE	0.870	0.633	0.60
willful disobedience	CA	0.870	0.625	0.60
willful violation	FL	0.843	0.645	0.45
enforce by contempt	TX	0.865	0.624	0.50
disobedience	NV	0.863	0.620	0.55
sanctions	ALL	0.915	0.583	0.60
willful violation	AL	0.825	0.643	0.70
financial misconduct	OH	0.840	0.628	0.80
breach of fid. duty	CA	0.853	0.619	0.85
willfully failed	PA	0.850	0.616	0.50

Table 4: Top 10 leverage words by combined leverage × PCA vulnerability. "Willful disobedience" — the term triggering contempt proceedings — is the most vulnerable high-leverage term. PCA Dmg = mean absolute similarity change at 16d.

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 5: PCA variance explained: legal corpus vs. Wikipedia control, both embedded with nomic-embed-text. The legal–Wikipedia gap is 2.8pp at 16 dimensions, too small to confirm a domain-specific effect. The 50pp gap with E5-Mistral on MTEB is model-architectural.

5 Definition Provenance and Jurisdictional Anchoring

A legal embedding without source provenance is semantically underdetermined. When a model embeds the term "contempt," it produces a single vector — but "contempt" as defined by Bouvier's (emphasizing disobedience to court authority), Webster's (a general state of disregard), Black's (distinguishing civil from criminal), and Illinois statute 750 ILCS 5/607.5 (specific enforcement mechanism for allocation orders) are four operationally distinct concepts. Without metadata indicating which definition produced the embedding, a downstream system cannot determine which meaning the vector represents.

We propose that every legal term embedding be accompanied by a provenance record specifying the metadata fields in Table 6.

The definitional drift results from Experiment 1 (Table 2) demonstrate why provenance matters quantitatively: Bouvier's and Webster's definitions of the same term differ by up to 36.4% in embedding space. Without provenance, a retrieval system has no way to determine which definition it is searching against. The cross-jurisdiction drift results from Experiment 2 (Table 3) compound this: even within a single dictionary tradition, "material change" in Nebraska and "material change" in Alaska are 33.7% apart. The embedding alone cannot recover which jurisdiction's meaning it encodes.

This provenance requirement is not merely a metadata best-practice — it is a precondition for semantic determinacy. An embedding without provenance is an embedding of an ambiguous input, and no amount of dimensionality or model quality can resolve ambiguity that was baked in at embedding time.

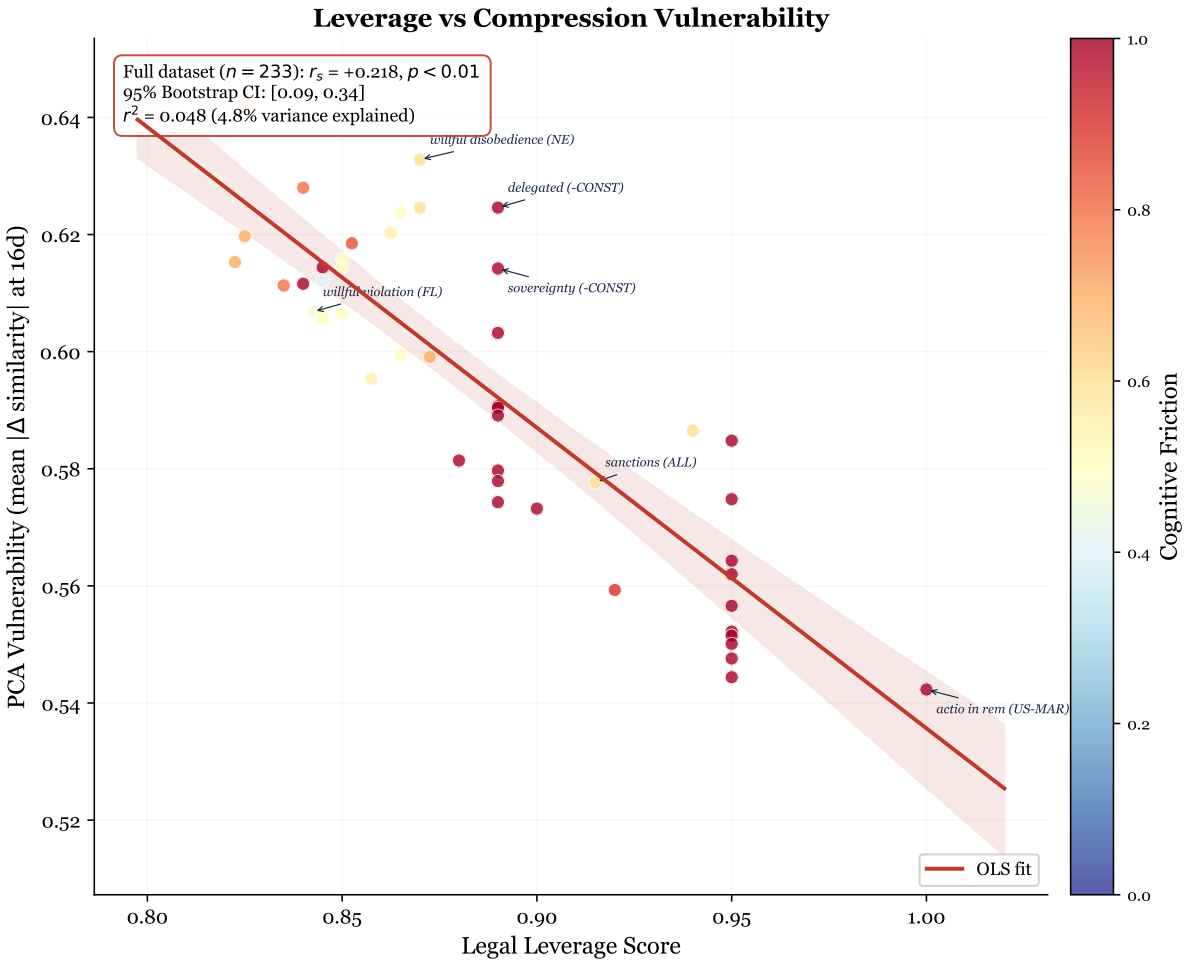

Figure 5: Leverage score vs. PCA vulnerability for 233 legal terms across 27 jurisdictions ($r_s = +0.22$, $p < 0.01$, 95% CI $[0.09, 0.34]$). Regression line with 95% confidence band shown.

6 Discussion

6.1 Dictionary Provenance as a First-Class Variable

The 23% mean definitional drift between Bouvier's and Webster's shows that dictionary choice is not a terminological convenience but a variable that determines embedding semantics. Bouvier's reflects 19th-century American common law, with definitions emphasizing historical precedent and property-based reasoning. Webster's strips legal specificity, producing definitions that are linguistically adequate but legally insufficient. Since embedding models inherit their implicit definitions from training data, the composition of that data — how much legal versus general text it contains — determines which dictionary's worldview the model encodes by default. No current benchmark measures this.

6.2 Cross-Jurisdiction Drift as a Hidden Failure Mode

The 33.7% drift between "material change" in Nebraska and Alaska is a practical problem for legal AI systems that serve multiple jurisdictions. A system trained on California case law will systematically under-rank Alaska precedent for the same legal concept, not because the concept differs, but because the surrounding statutory language produces different embeddings. This failure mode is invisible to standard benchmarks, which do not test cross-jurisdiction retrieval within a single legal domain. The 141-jurisdiction heatmap (Figure 4) suggests the problem is

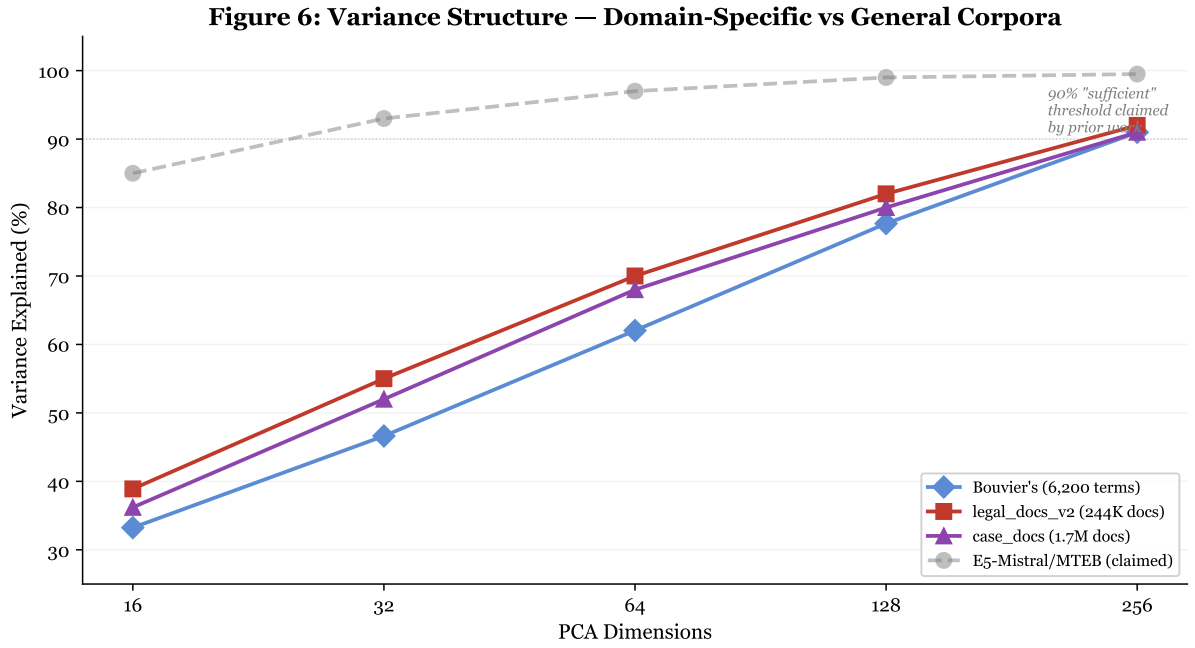

Figure 6: Variance explained by PCA across dimensionalities for four corpora embedded with nomic-embed-text and the E5-Mistral/MTEB reference. The 50pp gap is cross-model; within nomic-embed-text, legal and general text behave similarly (Table 5).

pervasive, not confined to a few edge cases.

6.3 Leverage–Vulnerability as a Risk Metric

The correlation between legal leverage and PCA vulnerability ($r_s = +0.22$) is modest but points toward a practical risk metric: before compressing embeddings for a legal application, compute the leverage scores of the terms most critical to that application and verify that their PCA vulnerability falls below an acceptable threshold. In our data, terms with leverage > 0.85 (the enforcement tier — contempt, sanctions, willful disobedience) have mean PCA damage of 0.62, meaning their similarity relationships change by 62% on average at 16 dimensions.

6.4 Structural Holes as Missing Protections

The clustering of cross-jurisdictional terms in Figure 8 visualizes a deeper problem: English has no word for *Pflichtteil* (children's compulsory inheritance share), *umgängessabotage* (contact sabotage by one parent), or *mu'sir* (one unable to pay, entitled to respite). These terms appear as isolated points with no nearby English-language neighbors. A retrieval system operating in English cannot find these concepts because they occupy structural holes in embedding space — regions where English legal vocabulary provides no coverage.

6.5 The Variance Gap Revisited

Our Wikipedia control experiment settles a question left open by our prior work [Thorarinson and Hensgen, 2026]: the 50pp gap between nomic-embed-text and E5-Mistral at 16 dimensions is architectural, not domain-specific. The legal–Wikipedia gap is only 2.8pp. This means claims about "legal text being harder to compress" based on aggregate variance are not supported; what *is* supported is that the compression damage, while model-level in aggregate, targets legally important terms disproportionately (Experiment 3). The distinction matters: the former would call for domain-specific models; the latter calls for domain-aware compression strategies.

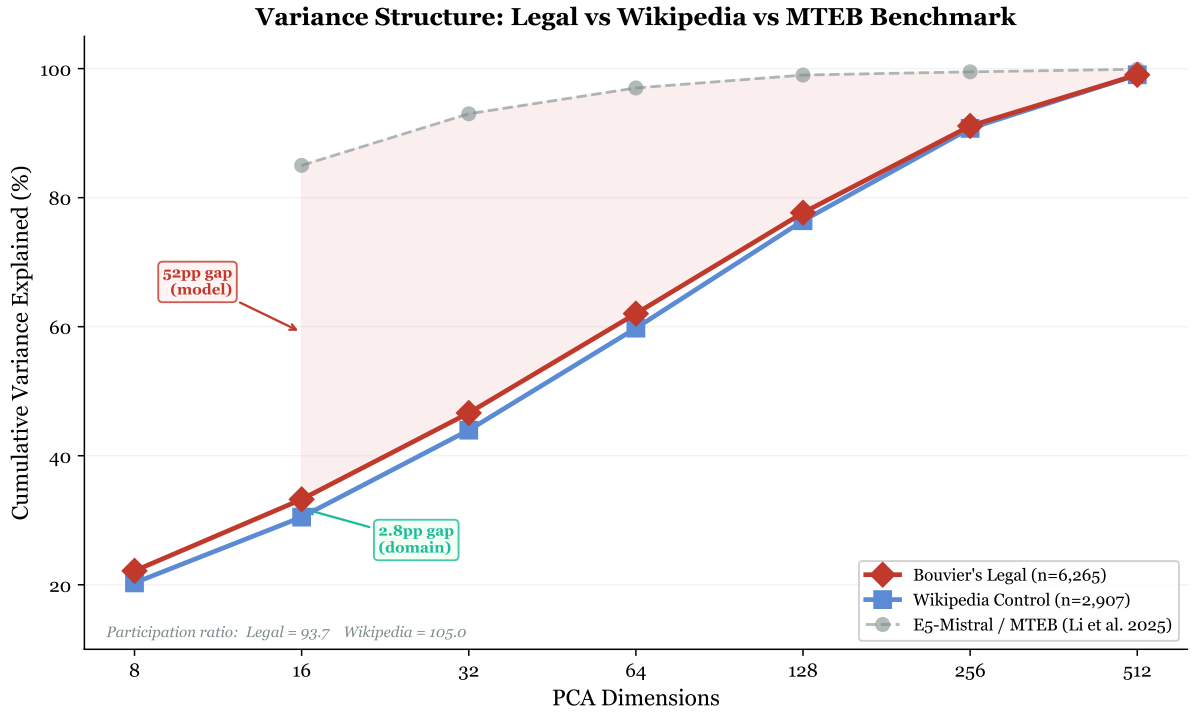

Figure 7: PCA variance explained: Bouvier's legal terms vs. Wikipedia control, both embedded with nomic-embed-text v1.5. The curves are nearly identical ($\leq$2.8pp gap), confirming the low variance at 16d is a model property. The E5-Mistral/MTEB reference (dashed) shows the cross-model gap.

6.6 Limitations

Our analysis uses a single embedding model (nomic-embed-text v1.5). Validation across architectures (BGE [Xiao et al., 2023], E5-Mistral [Wang et al., 2024], multilingual-e5) would determine whether definitional drift and the leverage–vulnerability correlation are model-specific or general. The Webster's control (46 terms) is small relative to Bouvier's (6,200); a larger general-dictionary dataset would strengthen the drift analysis. The leverage word database contains 233 terms across 27 jurisdictions; expansion to all 50 U.S. states (in progress) will improve statistical power. The cross-jurisdictional terms (19) are manually compiled; systematic extraction from multilingual legal databases [Niklaus et al., 2023b] would be valuable.

7 Conclusion

The same legal term, defined by different dictionaries, drifts by up to 36.4% in embedding space. The same term, applied in different jurisdictions, drifts by up to 33.7%. These are not edge cases; the mean drift across 44 overlapping terms is 23.0%, and the expanded 141-jurisdiction analysis shows drifts up to 41.9%. Dictionary provenance and jurisdictional context are determinants of embedding semantics, not optional metadata.

The terms most affected by PCA compression are not random. Legal leverage score correlates positively with compression vulnerability ($r_s = +0.22$, $p < 0.01$): the words that determine custody, imprisonment, and constitutional rights are disproportionately damaged. Meanwhile, a Wikipedia control confirms that the aggregate variance floor at 16 dimensions (33% for legal text, 30.5% for Wikipedia) is a property of the embedding model, not of legal language — but the non-uniform distribution of that damage across legally consequential terms is domain-specific and consequential.

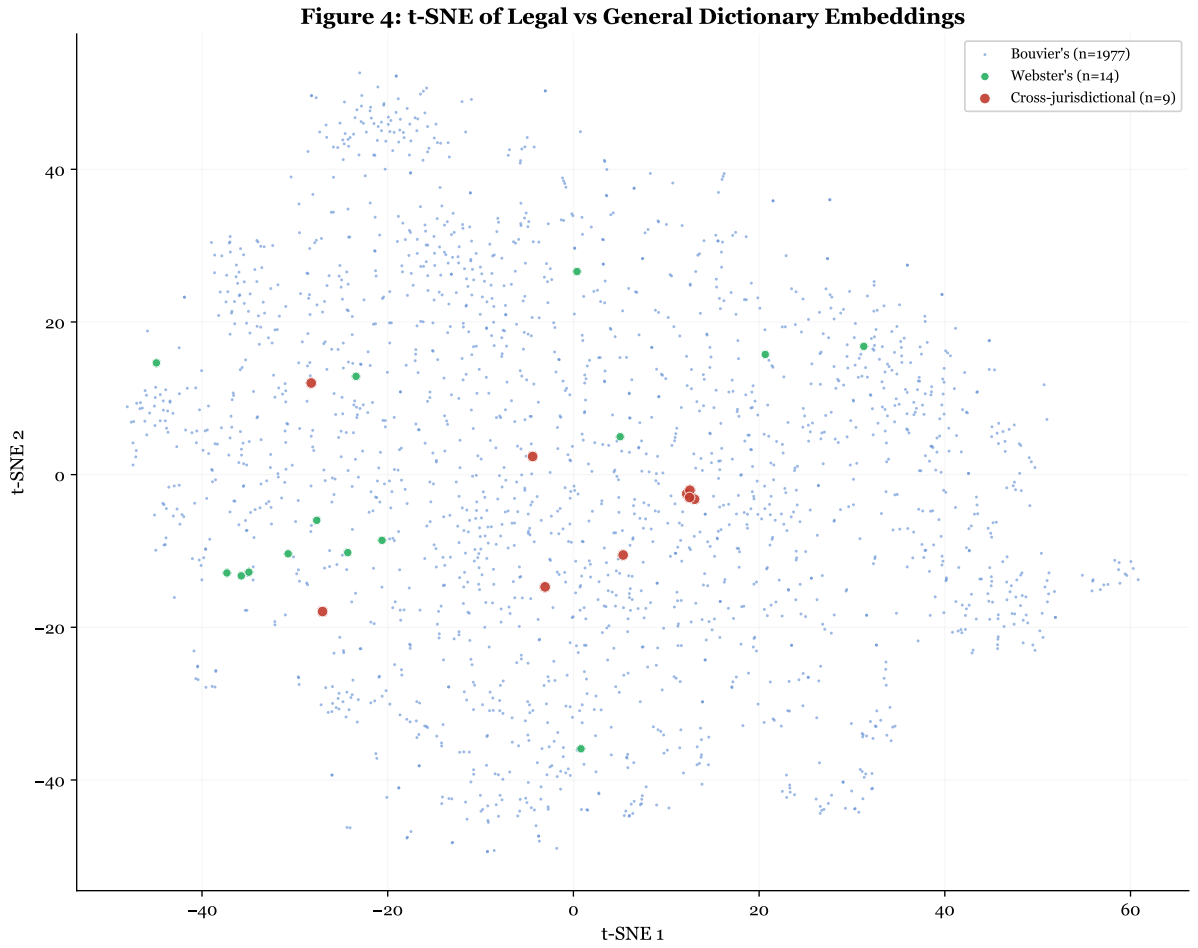

Figure 8: t-SNE visualization of the DictEmbed dataset. Bouvier's terms (blue, 2,000 sampled) fill the space. Webster's terms (green) scatter throughout the Bouvier's cloud at different specific locations. Cross-jurisdictional terms (red) cluster tightly in a distinct region — structural holes where English dictionaries provide no coverage.

These findings have practical implications: compression benchmarks validated on one model–corpus combination do not generalize, and the damage targets the terms that matter most. The embedding space is a map. Different dictionaries draw different maps. Compression erases the features that distinguish them — and the features it erases first are the ones that determine outcomes.

References

John Bouvier. *A Law Dictionary, Adapted to the Constitution and Laws of the United States of America*. Childs & Peterson, Philadelphia, 6 edition, 1856.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. *Findings of 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, Dublin, Ireland, 2022.

Metadata Field	Description
Term	The legal term as embedded
Definition source	The dictionary or statute from which the definition was drawn (e.g., Bouvier's 1856, Black's 11th ed., 750 ILCS 5/607.5)
Jurisdiction	The jurisdiction in which the definition has legal authority (e.g., Illinois, Federal, Islamic law)
Date / version	The edition or effective date of the source (e.g., 1856, 2024 supplement)
Procedural context	The procedural posture in which the term is operative (e.g., enforcement, modification, initial determination)
Authority level	The normative weight of the source (constitutional, statutory, regulatory, judicial, treatise, dictionary)
Consequence category	The domain of legal consequence triggered by the term (liberty, custody, property, procedural, administrative)

Table 6: Provenance metadata for legal term embeddings. Embedding a legal term without these fields risks collapsing multiple operationally distinct meanings into a single vector.

Junyun Cui, Xiaoyu Shen, Fei Wen, et al. A survey on legal judgment prediction. *arXiv preprint arXiv:2204.04859*, 2023. URL https://arxiv.org/abs/2204.04859.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics, 2019.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, 2019. URL https://aclanthology.org/D19-1006/.

John Rupert Firth. A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford, 1957.

Stanisław Goźdź-Roszkowski and Gianluca Pontrandolfo. *Phraseology in Legal and Institutional Settings: A Corpus-based Interdisciplinary Perspective*. Law, Language and Communication. Routledge, 2020.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2308.11462.

Zellig S. Harris. Distributional structure. *Word*, 10(2–3):146–162, 1954.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An expert-annotated NLP dataset for legal contract review. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2103.06268.

Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, 2 edition, 2002.

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.

David Mellinkoff. *The Language of the Law*. Little, Brown and Co., Boston, 1963.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. SemEval-2022 Task 1: CODWOE – comparing dictionaries and word embeddings. In *SemEval-2022*, 2022. URL https://arxiv.org/abs/2205.13858.

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.

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.

Joel Niklaus, Veton Matoshi, Pooja Rani, et al. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of EMNLP 2023*, 2023a. URL https://arxiv.org/abs/2301.13126.

Joel Niklaus, Veton Matoshi, et al. MultiLegalPile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*, 2023b. URL https://arxiv.org/abs/2306.02069.

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 Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014.

Nicholas Pipitone et al. LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*, 2024. URL https://arxiv.org/abs/2408.10343.

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

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.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

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*, 2025. URL https://arxiv.org/abs/2508.17744.

Joel Thorarinson and Allison Hensgen. The dimensionality illusion: Why PCA variance does not equal semantic information in text embeddings. *arXiv preprint*, 2026. Paper 1 in this series.

Peter M. Tiersma. *Legal Language*. University of Chicago Press, Chicago, 1999.

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.

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.

Konrad Zweigert and Hein Kötz. *An Introduction to Comparative Law*. Oxford University Press, Oxford, 3 edition, 1998.