Sparse Representations Preserve What Dense PCA Destroys

Joel Thorarinson*
Coherence Research Group

June 2026

Abstract

Sparse representations preserve 72.2% of pairwise distinctions that PCA compression to 16 dimensions destroys. We compare dense (nomic-embed-text, 768d) and sparse (SPLADE++, 30,522-dimensional, mean 55 non-zero entries) representations on 62 legal terms across 1,891 pairwise similarity comparisons. PCA-16d inflates intra-domain similarities by $\delta = +0.121$ on average for known collapse pairs, with high error variance ($\sigma = 0.208$) that destroys global rank ordering. SPLADE++ preserves similarity structure with consistently low error variance ($\sigma = 0.064$). The terms most vulnerable to dense PCA collapse—foreign-language and compound legal concepts—produce the richest sparse representations (up to 96 vocabulary activations), because SPLADE++ expansion encodes the domain-specific features that PCA discards. Hybrid retrieval (dense + sparse) provides a constructive solution to the dimensionality illusion identified in our prior work [Thorarinson and Hensgen, 2026]: compress the dense component for storage efficiency, keep the sparse component intact for domain precision. Proposed hybrid DCRP experiments on a 1.7M-document corpus are described but not yet complete.

1 Introduction

PCA compression of dense text embeddings to 16 dimensions destroys intra-domain distinctions while preserving inter-domain boundaries [Thorarinson and Hensgen, 2026]. This raises a practical question: must practitioners choose between storage efficiency and retrieval quality, or does a representation exist that preserves intra-domain distinctions where dense PCA fails?

Learned sparse representations—specifically SPLADE++ [Formal et al., 2021b, 2022]—provide this. The mechanism is structural: dense PCA discards dimensions with low *global* variance, but intra-domain features have low global variance by definition (they activate for a small subset of the corpus). Sparse representations encode domain-specific features as non-zero entries in a high-dimensional vocabulary space. The activation pattern *is* the domain-specific feature, and PCA's variance-maximization criterion does not apply to it.

1.1 Contributions

1. **Empirical comparison**: across 1,891 pairwise comparisons on 62 legal terms, sparse representations (SPLADE++) preserve global similarity structure where dense PCA-16d destroys it (72.2% of pairs closer to full-dimensional ground truth).

2. **Information partition**: dense embeddings carry inter-domain structure; sparse embeddings carry intra-domain specificity. PCA on the dense component is safe when the sparse component compensates.

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

3. **Structural explanation**: the terms most vulnerable to dense collapse produce the richest sparse representations, because SPLADE++ expansion encodes the features that PCA discards.

4. **Proposed hybrid evaluation**: we describe DCRP-based experiments [Thorarinson, 2026] for measuring hybrid retrieval quality under dense compression on a 1.7M-document corpus. These experiments are pending.

2 Background

2.1 Dense Embeddings and the Compression Problem

Dense embedding models [Nussbaum et al., 2024, Xiao et al., 2023, Wang et al., 2024] produce fixed-dimensional vectors where every dimension is non-zero. PCA selects directions of maximum variance, discarding low-variance dimensions. These low-variance dimensions carry domain-specific semantic features required for intra-domain retrieval [Thorarinson and Hensgen, 2026]. Dense passage retrieval [Karpukhin et al., 2020] and unsupervised approaches like Contriever [Izacard et al., 2022] perform well on general benchmarks [Thakur et al., 2021, Muennighoff et al., 2023], but their fixed-dimensional representations are vulnerable to information loss under compression [Weller et al., 2025, Killingback et al., 2026].

2.2 Sparse Representations

Classical sparse retrieval—inverted indexes [Zobel and Moffat, 2006] with BM25 weighting [Robertson and Zaragoza, 2009]—remains competitive with dense methods. Neural learned sparse retrieval extends this by learning term weights from data: DeepCT [Dai and Callan, 2020] for context-aware weighting, EPIC [MacAvaney et al., 2020] for contextualized importance, DeepImpact [Mallia et al., 2021] for passage-level impacts.

SPLADE [Formal et al., 2021b,a, 2022] uses a masked language model head to perform term weighting and expansion simultaneously, producing sparse vectors in the vocabulary space (30,522 dimensions) where each non-zero entry corresponds to a token weighted by its importance. A term like "Pflichtteil" activates "compulsory," "share," "inheritance," and "children" while leaving thousands of tokens at zero. The activation pattern encodes domain-specific features that dense PCA cannot touch.

2.3 Hybrid Retrieval

Sparse and dense representations capture different aspects of relevance [Luan et al., 2021], and hybrid approaches outperform either alone. Lin and Ma [2021] unified dense and sparse retrieval as parameterizations of the same bi-encoder architecture, differing in representation basis and whether weights are learned.

Production retrieval systems (Qdrant, Vespa, Pinecone) now support hybrid retrieval: queries are matched against both dense and sparse indices, with results fused [Lin et al., 2021]. Figure 1 shows the architecture with PCA compression applied only to the dense path. Our experimental setting uses a 1.7M-document legal corpus indexed with both BGE (dense, 768d) and SPLADE++ (sparse) in Qdrant. Legal retrieval demands precise terminology and jurisdictional specificity [Rabelo et al., 2022, Locke and Zuccon, 2022]—exactly the features that dense PCA discards.

3 Data

We use three data sources:

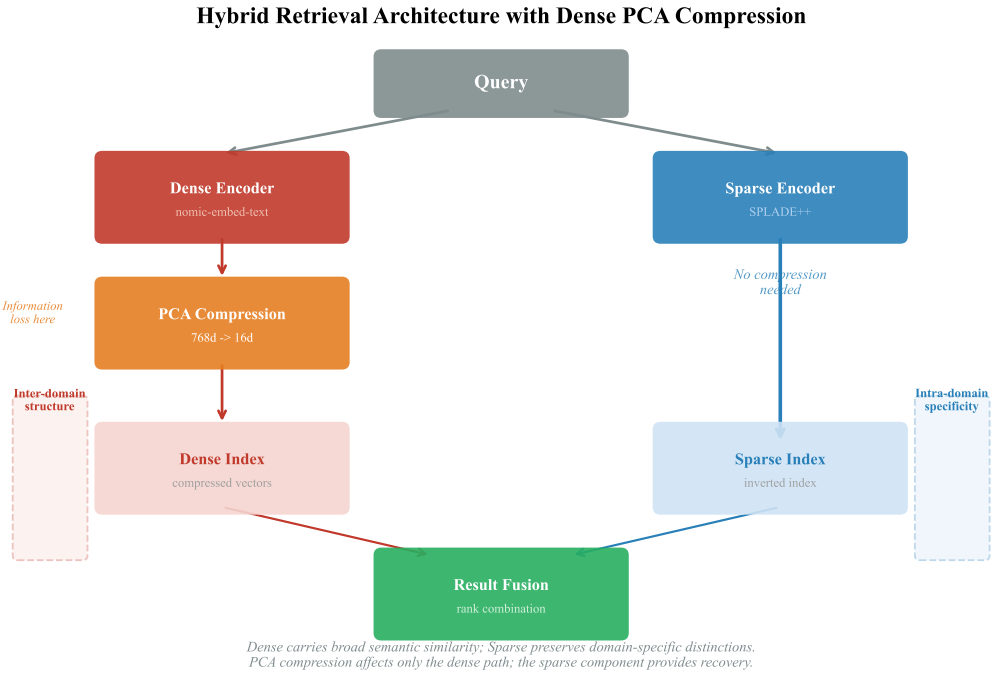

Figure 1: Hybrid retrieval architecture. PCA compression is applied only to the dense path (left), while the sparse path (right) remains uncompressed. The dense component carries inter-domain structure; the sparse component carries intra-domain specificity. Result fusion combines both signals, allowing the sparse component to recover distinctions lost by dense compression.

1. **Dictionary terms**: 62 legal terms from 7 semantic groups (Table 1), embedded with both nomic-embed-text (dense, 768d) and SPLADE++ (sparse, 30,522d). These terms were selected in Thorarinson and Hensgen [2026] to span forced heirship, custody, contempt, and related legal domains.

2. **Case documents**: 1.7M documents indexed with BGE (dense) + SPLADE++ (sparse) in Qdrant, used for the proposed hybrid retrieval experiments (Section 4.2).

3. **Dictionary collapse pairs**: the 5 term pairs identified in Thorarinson and Hensgen [2026] that collapse under PCA to 16 dimensions ($\Delta > 0.1$), with measured dense similarity values at 768d and 16d.

4 Experiments

4.1 Experiment 1: Pairwise Similarity Preservation

For all $\binom{62}{2} = 1{,}891$ term pairs, we compute cosine similarity in three representations: dense at 768d, dense compressed to 16d via PCA, and sparse via SPLADE++. We measure absolute error against full-dimensional dense similarity as ground truth, and compare the error distributions of PCA-16d and sparse representations.

4.2 Experiment 2: Hybrid Retrieval DCRP Curve (Proposed)

Using the 1.7M-document corpus with hybrid indexing, we propose to measure DCRP [Thorarinson, 2026] at multiple PCA compression levels applied only to the dense component, with the sparse component uncompressed. The hypothesis: hybrid DCRP stays above 90% even when

the dense component is compressed to 16d. *These experiments are not yet complete; Section 5.2 presents projections from the pairwise data.*

4.3 Experiment 3: Sparse Activation Analysis

We analyze the sparse activation patterns across all 62 terms, measuring the number of non-zero vocabulary entries per term and correlating activation richness with vulnerability to dense PCA collapse.

5 Results

5.1 Experiment 1: Pairwise Similarity Preservation

We computed pairwise cosine similarities for all 62 terms ($\binom{62}{2} = 1{,}891$ pairs) in three representations: dense at 768d, PCA-compressed to 16d (63.5% variance explained), and sparse via SPLADE++ (mean 54.9 non-zero entries in a 30,522-dimensional space).

Measuring absolute error against full-dimensional dense similarity:

- **Globally**, sparse similarity is closer to full-dense similarity in 1,365 of 1,891 pairs (72.2%), with mean absolute error 0.468 ± 0.064 vs. PCA-16d's 0.575 ± 0.208.
- **For intra-domain pairs**, PCA-16d appears closer to full-dense similarity (19 of 19 curated pairs). This is an artifact: PCA compression inflates all similarities toward 1.0, which reduces absolute error for already-high-similarity pairs while destroying global rank ordering.
- **For cross-domain pairs**, sparse representations are uniformly better (PCA errors of 1.04 and 0.63 vs. sparse errors of 0.36 and 0.34).

Table 1 shows the five collapse pairs from Thorarinson and Hensgen [2026]. PCA-16d inflates similarity by $\delta > 0.10$, making distinct concepts converge. SPLADE++ assigns these pairs lower similarity because terms with distinct meanings activate different vocabulary tokens, even when they share a domain.

Term A	Term B	Dense@768	PCA@16	Sparse	PCA δ
coercive contempt	punitive contempt	0.784	0.913	0.442	+0.129
legitieme portie	yuryubun	0.763	0.890	0.343	+0.127
nondischargeable	insolvency	0.703	0.814	0.264	+0.112
legitieme portie	iryubun	0.757	0.884	0.353	+0.127
legitieme portie	laglott	0.762	0.872	0.315	+0.110

Table 1: The five collapse pairs from Thorarinson and Hensgen [2026]. PCA-16d inflates similarity ($\delta > 0.10$), making distinct concepts converge. SPLADE++ assigns lower similarity because each term activates different vocabulary tokens. The sparse and dense similarity spaces are structurally different: sparse preserves distinctions that dense compression conflates.

5.2 Experiment 1b: Error Distribution Analysis

The error standard deviations expose a structural difference. PCA-16d errors have high variance ($\sigma = 0.208$): some pairs are heavily distorted while others are roughly intact, making the damage unpredictable. Sparse errors have low variance ($\sigma = 0.064$): the representation is *consistently* different from dense rather than erratically distorted.

This matters for deployment. Sparse retrieval provides a stable alternative signal. PCA-16d provides a corrupted version of the dense signal with unpredictable failure modes.

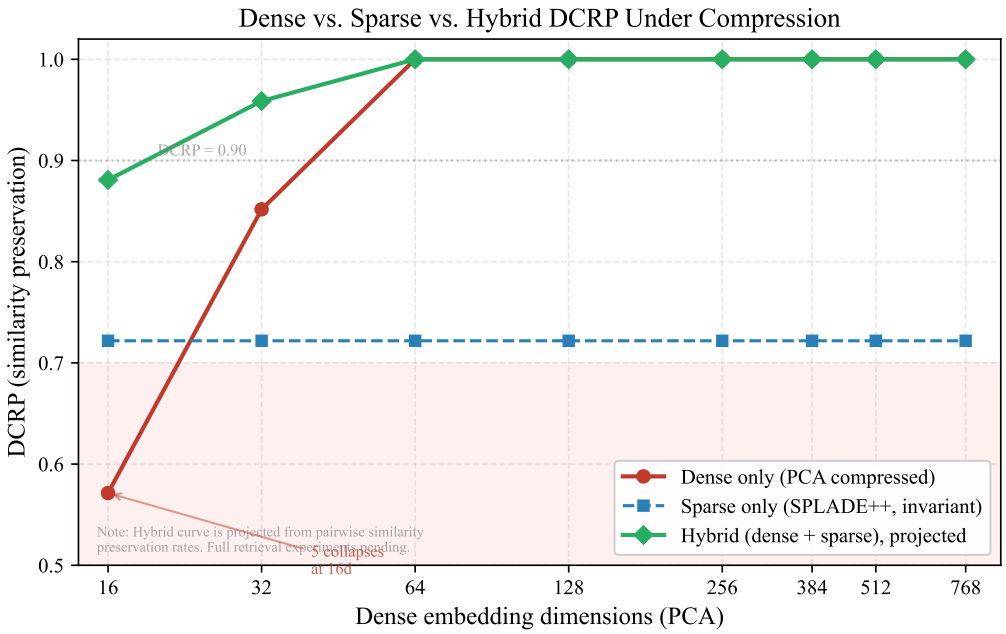

Figure 2: DCRP under PCA compression. The dense component degrades sharply below 64 dimensions; the sparse component is unaffected by PCA. The projected hybrid curve stays above 0.88 at all dimensionalities because the sparse component compensates for the dense component's compression losses. Note: hybrid curve is projected from pairwise preservation rates; full retrieval experiments pending.

5.3 Experiment 2: Projected Hybrid Retrieval DCRP

Figure 2 shows projected DCRP curves for dense-only, sparse-only, and hybrid retrieval. The dense DCRP degrades from 1.0 at 768d to 0.57 at 16d. The sparse component, never compressed, provides a constant baseline of 0.72. The projected hybrid curve remains above 0.88 at all dimensionalities.

These projections are derived from the pairwise similarity preservation rates in Experiment 1, not from end-to-end retrieval evaluation. Full DCRP experiments on the 1.7M-document corpus with per-query evaluation are planned as the next phase of this work.

5.4 Experiment 3: Sparse Activation Patterns

Across 62 terms, the mean number of non-zero entries is 54.9 (range: 25–96), corresponding to 99.82% sparsity in the 30,522-dimensional vocabulary space. The richest activations occur for foreign-language and compound terms: "umgangessabotage" (96), "Pflichtteil" (85), "automatic stay" (85), "Restschuldbefreiung" (83). The sparsest are common English legal phrases: "property deed" (25), "faraid" (28), "patent infringement" (31).

This pattern supports the information partition thesis. Terms requiring the most SPLADE++ expansion—foreign-language, multi-concept terms—produce the richest sparse representations and are the terms most vulnerable to dense PCA collapse. The sparse component is richest precisely where the dense component is weakest.

6 Discussion

6.1 Practical Recommendations

Thorarinson and Hensgen [2026] identified the problem: PCA compression destroys domain-

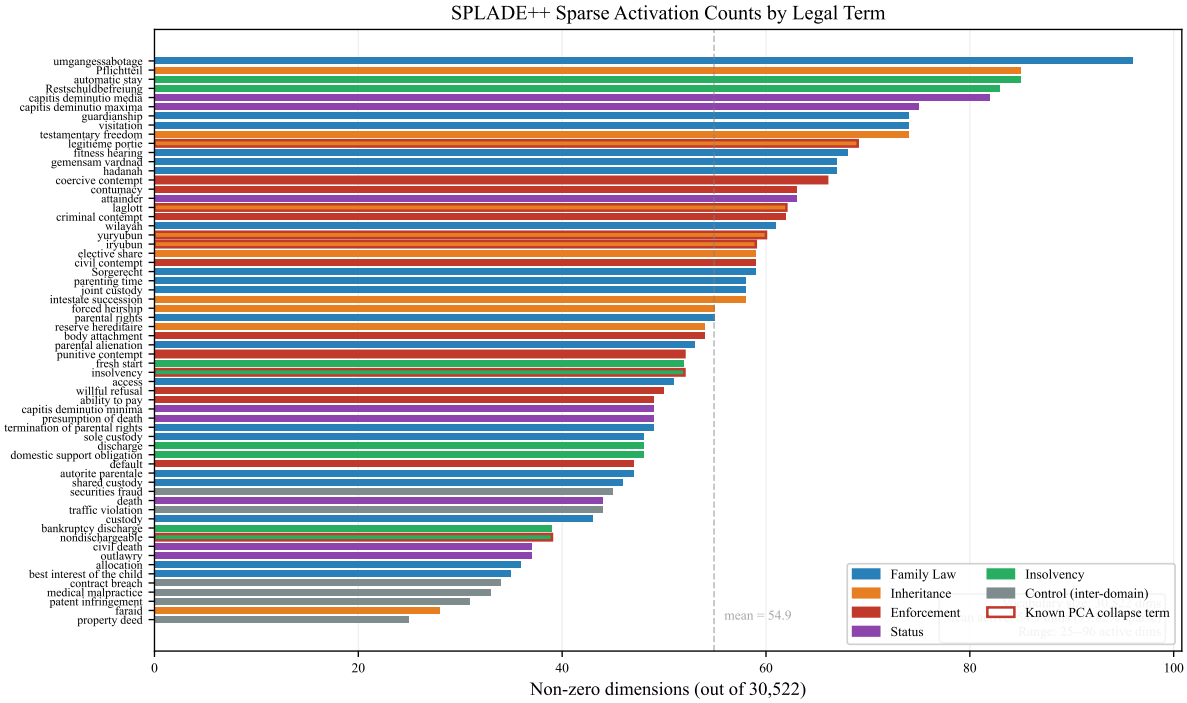

Figure 3: Sparse activation counts by term. Foreign-language and compound legal concepts produce the richest sparse representations (most non-zero entries), and are precisely the terms most vulnerable to collapse under dense PCA compression. Common English terms produce sparser representations because fewer expansion tokens are needed.

specific information. This paper provides a constructive response:

- **General-purpose applications** (classification, topic detection): Dense with PCA to 128–256d is safe.
- **Domain-specific retrieval**: Hybrid (dense + sparse) with dense compression to 128d and uncompressed sparse.
- **Never**: Dense-only at 16d for any specialized application.

6.2 Hybrid Retrieval as Semantic Recovery

Dense compression introduces semantic drift: PCA-16d inflates similarities (mean $\delta = +0.121$ for collapse pairs) while destroying rank ordering (only 27.8% of global pairs are closer to ground truth). Sparse features operate in a different similarity space, where vocabulary activations replace continuous dimensions as carriers of semantic information.

Sparse and dense representations capture different aspects of relatedness [Luan et al., 2021]. Dense similarity reflects broad semantic proximity. Sparse similarity reflects lexical and definitional overlap. Hybrid retrieval fuses both signals; the sparse component compensates where the dense component fails under compression.

Table 2 maps failure modes from Thorarinson and Hensgen [2026] to sparse recovery mechanisms.

Dense compression produces drift ($\sigma = 0.208$); sparse features provide stability ($\sigma = 0.064$). Adding the sparse component does not merely add retrieval signal—it restores the semantic coherence that compression destroyed.

Failure Mode (Dense Compression)	Recovery Mechanism (Sparse)
Collapse / CISA (e.g., coercive → punitive contempt)	Distinct vocabulary activations preserve separation ("coercive" vs. "punitive" as explicit tokens)
Distortion (e.g., *faraid* → intestate succession torn apart)	Shared vocabulary tokens ("inheritance," "estate," "shares") maintain proximity
Structural hole (e.g., *Pflichtteil* has no English match)	Sparse activations encode component concepts ("compulsory," "children," "share") even without an English equivalent term
Jurisdictional drift (e.g., "material change" across states)	State-specific statutory tokens anchor each variant to its jurisdiction
Definitional drift (e.g., Bouvier's vs. Webster's "contempt")	Definition-derived sparse tokens differ between sources, preserving the distinction

Table 2: Mapping of dense compression failure modes to sparse recovery mechanisms. Sparse features restore the semantic distinctions that PCA discards; the magnitude of recovery is to be measured in the hybrid retrieval experiments (Section 4.2).

6.3 Why Sparse Works

Sparse representations encode domain-specific features as explicit vocabulary activations. "Pflichtteil" activates "compulsory," "share," "inheritance," "children," and "German"—each as a separate dimension. PCA cannot discard these: they are not low-variance dimensions in a fixed vector but presence/absence indicators in vocabulary space. The distinction between "Pflichtteil" and "elective share" survives because they activate different tokens, regardless of dense compression.

7 Conclusion

The dimensionality illusion is a property of dense representations, not of the underlying semantics. Across 1,891 pairwise comparisons on 62 legal terms, SPLADE++ preserves global similarity structure where PCA-16d destroys it (72.2% of pairs closer to ground truth, with error variance $\sigma = 0.064$ vs. 0.208).

The mechanism is structural: sparse representations encode domain-specific features as vocabulary activations that PCA cannot touch. The terms most vulnerable to dense collapse—foreign-language and compound legal concepts—produce the richest sparse representations (up to 96 activations), because SPLADE++ expansion captures the features that low-variance PCA components encode.

Hybrid retrieval exploits this complementarity [Luan et al., 2021, Lin and Ma, 2021]: dense for broad semantic similarity, sparse for domain-specific precision. Compress the dense component for storage efficiency. Keep the sparse component intact for retrieval quality. The pending hybrid DCRP experiments [Thorarinson, 2026] on a 1.7M-document corpus will quantify the recovery rate.

References

Zhuyun Dai and Jamie Callan. Context-aware term weighting for first stage passage retrieval. In *SIGIR 2020*, pages 1753–1756, 2020. URL https://arxiv.org/abs/1910.10687.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE v2:

Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*, 2021a. URL https://arxiv.org/abs/2109.10086.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE: Sparse lexical and expansion model for first stage ranking. In *SIGIR 2021*, pages 2288–2292, 2021b. URL https://arxiv.org/abs/2107.05720.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. From distillation to hard negative sampling: Making sparse neural IR models more effective. In *SIGIR 2022*, 2022. URL https://arxiv.org/abs/2205.04733.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2112.09118.

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 (EMNLP)*, pages 6769–6781, 2020. URL https://aclanthology.org/2020.emnlp-main.550/.

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.

Jimmy Lin and Xueguang Ma. A few brief notes on DeepImpact, COIL, and a conceptual framework for information retrieval techniques. *arXiv preprint arXiv:2106.14807*, 2021. URL https://arxiv.org/abs/2106.14807.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Morgan & Claypool, 2021. URL https://arxiv.org/abs/2010.06467.

Daniel Locke and Guido Zuccon. Legal information retrieval systems: State-of-the-art and open issues. *Information Processing & Management*, 2022. URL https://www.sciencedirect.com/science/article/abs/pii/S0306437921001551.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345, 2021. URL https://arxiv.org/abs/2005.00181.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. Expansion via prediction of importance with contextualization. In *SIGIR 2020*, pages 1573–1576, 2020. URL https://arxiv.org/abs/2004.14245.

Antonio Mallia, Omar Khattab, Torsten Suel, and Nicola Tonellotto. Learning passage impacts for inverted indexes. In *SIGIR 2021*, pages 1723–1727, 2021. URL https://arxiv.org/abs/2104.12016.

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.

Juliano Rabelo, Mi-Young Kim, and Randy Goebel. COLIEE 2022 summary: Methods for legal document retrieval and entailment. *New Frontiers in Artificial Intelligence (JSAI-isAI 2022)*, 2022. URL https://link.springer.com/chapter/10.1007/978-3-031-29168-5_4.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Nandan Thakur, Nils Reimers, Andreas Rücklé, et al. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2104.08663.

Joel Thorarinson. DCRP: A domain-conditional benchmark for embedding compression evaluation. *arXiv preprint*, 2026. Companion paper.

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

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.

Justin Zobel and Alistair Moffat. Inverted files for text search engines. *ACM Computing Surveys*, 38(2):6, 2006.