DCRP: A Domain-Conditional Benchmark for Embedding Compression Evaluation

Joel Thorarinson*
Coherence Research Group

June 2026

Abstract

Standard embedding benchmarks evaluate retrieval and classification across domains but not *within* domains. Embedding compression methods—PCA, quantization, Matryoshka truncation—are therefore validated on inter-domain discrimination while intra-domain precision goes unmeasured. We introduce **DCRP** (Domain-Conditional Retrieval Precision), a metric and benchmark framework that evaluates compression on the task it degrades most: fine-grained retrieval within specialized corpora. We present pilot results from a 62-term legal dictionary experiment: at PCA to 16 dimensions, inter-domain retrieval metrics show zero degradation while DCRP detects 5 intra-domain term collapses, including the coercive/punitive contempt distinction that determines jury trial rights. These results demonstrate that MTEB-style scores provide no information about domain-specific compression quality in the low-dimensionality regime. We define the DCRP metric formally, describe the framework for constructing domain-specific query sets with consequence weighting, and outline the path toward a full benchmark release with human relevance judgments.

1 Introduction

MTEB [Muennighoff et al., 2023] and BEIR [Thakur et al., 2021] have become the standard evaluation suites for text embedding models. Models from Sentence-BERT [Reimers and Gurevych, 2019] through DPR [Karpukhin et al., 2020], BGE [Xiao et al., 2023], E5 [Wang et al., 2024], and nomic-embed-text [Nussbaum et al., 2024] are compared on these benchmarks as a matter of course.

Both suites are predominantly *inter-domain*. A retrieval task asks whether the system can find relevant documents from a mixed-domain corpus. A classification task asks whether it can assign documents to broad categories. These tasks exercise the principal components of the embedding space—the high-variance dimensions encoding topic, language, and register. Domain-specific benchmarks have begun to address this gap: FinMTEB [Tang and Yang, 2025] showed that general-purpose embeddings underperform domain-adapted models on financial texts, and ChemTEB [Kasmaee et al., 2024] documented analogous failures in chemistry.

What remains untested is *intra-domain* precision under compression. Can a compressed embedding distinguish German *Pflichtteil* from French *réserve héréditaire*? Can it separate "ability to pay" from "willful refusal" in child support contempt proceedings? These tasks matter for legal retrieval [Goebel et al., 2024, Guha et al., 2023, Pipitone et al., 2024], medical search [Tsatsaronis et al., 2015, Roberts et al., 2021], and financial analysis [Chen et al., 2021, Tang and Yang, 2025]—and they exercise the long-tail dimensions that PCA discards first.

Our prior work [Thorarinson and Hensgen, 2026] showed that PCA compression to 16 dimensions destroys these dimensions while preserving the principal components, producing a

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

pattern where MTEB scores remain high while intra-domain retrieval quality collapses. DCRP is designed to measure the collapse that MTEB misses.

1.1 Contributions

1. We define **DCRP** (Domain-Conditional Retrieval Precision), a metric that measures retrieval quality within semantic domains rather than across them, with an optional consequence-weighting scheme for high-stakes applications.

2. We present **pilot results from DCRP-Legal**, using 62 legal terms from 7 semantic groups and 6,200 Bouvier's Dictionary entries, demonstrating the metric's sensitivity to compression damage that inter-domain metrics miss.

3. We show that at 16 dimensions, **MTEB-style inter-domain metrics report zero degradation** while DCRP detects 5 intra-domain collapses—confirming that the two metric families are decorrelated in the low-dimensionality regime.

4. We describe a **framework for constructing domain-specific DCRP benchmarks** with human relevance judgments, applicable to legal, medical, financial, and chemical domains.

2 Related Work

2.1 Embedding Benchmarks and Their Limitations

MTEB [Muennighoff et al., 2023] evaluates embeddings on 56 datasets across 8 task categories; BEIR [Thakur et al., 2021] provides zero-shot retrieval evaluation across 18 datasets. Both primarily test inter-domain discrimination. Domain-specific benchmarks—FinMTEB [Tang and Yang, 2025] for finance, ChemTEB [Kasmaee et al., 2024] for chemistry, LegalBench-RAG [Pipitone et al., 2024] for legal RAG—test model quality within specialized domains. None of these evaluate *compression sensitivity*: how retrieval quality degrades under dimensionality reduction within a domain.

2.2 Embedding Compression Methods

Matryoshka Representation Learning [MRL; Kusupati et al., 2022] trains embeddings for graceful truncation, with extensions including Matryoshka-Adaptor [Yoon et al., 2024] and SMEC [Zhang et al., 2025]. PCA-based compression [Raunak, 2017, Zhang et al., 2024, Bengtsson, 2025] and quantization [Huerga-Perez et al., 2025] are widely deployed. Scaling law analysis [Killingback et al., 2026] suggests that required embedding dimension grows with corpus specificity. All these methods are evaluated on MTEB or BEIR—benchmarks that, as we show, are insensitive to the failures compression introduces in specialized domains.

2.3 Domain-Specific Retrieval

Legal retrieval is evaluated through COLIEE [Goebel et al., 2024] and LegalBench [Guha et al., 2023]; medical retrieval through BioASQ [Tsatsaronis et al., 2015] and the TREC Clinical Trials track [Roberts et al., 2021]; financial retrieval through FinQA [Chen et al., 2021] and FinMTEB [Tang and Yang, 2025]. These benchmarks demonstrate that domain-specific evaluation reveals failures invisible to general benchmarks. DCRP extends this principle to compression sensitivity.

2.4 IR Evaluation Metrics

Standard retrieval metrics—Precision@k, Recall@k, NDCG [Järvelin and Kekäläinen, 2002]—are surveyed in Voorhees and Harman [2005] and Manning et al. [2008]. DCRP builds on Precision@k but conditions on domain membership, producing a metric sensitive to intra-domain discrimination loss under compression.

3 The DCRP Metric

3.1 Definition

For a query q in domain $\mathcal{D}$, let $\text{rel}(q) \subset \mathcal{D}$ be the set of relevant documents. DCRP@k averages the fraction of top-k retrieved results that are relevant, across all queries targeting a given domain [Thorarinson and Hensgen, 2026, Section 4.5]:

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

The key design choice is the denominator: normalizing by $\min(k, |\text{rel}(q)|)$ ensures that queries with fewer than k relevant documents are not penalized for the corpus's sparsity in that sub-domain.

3.2 Design Principles

DCRP differs from standard retrieval metrics in three respects:

1. **Domain-conditional**: both queries and relevance judgments are defined *within* a single domain. The retrieval corpus for each query is restricted to same-domain documents.

2. **Compression-aware**: each query set is evaluated at multiple compression levels (16, 32, 64, 128, 256, 384, 512, and full dimensionality), producing a degradation curve rather than a single score.

3. **Consequence-weighted**: queries can be weighted by the practical severity of retrieval failure, distinguishing "inconvenient miss" from "catastrophic miss."

3.3 Consequence-Weighted DCRP

Not all retrieval failures are equal. Failing to retrieve the correct definition of "easement" in a real estate context is inconvenient; failing to distinguish "coercive contempt" from "punitive contempt" can determine whether a person receives a jury trial. To capture this asymmetry, we define Consequence-Weighted DCRP (CW-DCRP):

$$\text{CW-DCRP@}k = \frac{1}{|Q_{\mathcal{D}}|} \sum_{q \in Q_{\mathcal{D}}} \text{DCRP@}k(q) \cdot L(q) \tag{2}$$

where $L(q)$ is a legal consequence weight assigned to query q based on the severity of retrieval failure:

CW-DCRP inherits all properties of DCRP but amplifies the signal from high-stakes queries. In our dictionary collapse experiments, the coercive/punitive contempt pair ($L = 4$, jury trial right) and joint/sole custody pair ($L = 3$, parental rights) are the pairs with the highest collapse under PCA to 16 dimensions—meaning unweighted DCRP underestimates the practical damage of compression in the regime where it matters most.

$L(q)$	Level	Description
1	Low	Administrative or procedural inconvenience; no rights affected
2	Medium	Financial consequence; incorrect damages calculation, missed deadline
3	High	Liberty or custody consequence; contempt classification, parental rights
4	Catastrophic	Constitutional rights affected; jury trial lost, wrongful incarceration

Table 1: Legal consequence weight levels for CW-DCRP. The weight $L(q)$ scales retrieval precision by the severity of failure, ensuring that compression damage to high-consequence terms is penalized proportionally.

4 DCRP-Legal: The Legal Domain Benchmark

4.1 Corpus

DCRP-Legal uses four sub-corpora:

Sub-corpus	Size	Source
Bouvier's terms	6,200	Public domain (1856)
Leverage words	233	27 jurisdictions, scored
Cross-jurisdictional	19	7 legal traditions
Control (Webster's)	46	General-language

Table 2: DCRP-Legal corpus composition.

4.2 Planned Query Sets

The pilot experiments in this paper use term-level dictionary collapse (Section 5) rather than full query-based retrieval evaluation. The full DCRP-Legal benchmark will include 250 intra-domain queries with human relevance judgments across five categories:

- **Term disambiguation** (50 queries): distinguish "custody" from "guardianship" from "allocation"—terms that occupy nearby regions of embedding space but carry distinct legal consequences.
- **Cross-jurisdiction matching** (50 queries): find the equivalent term in another jurisdiction (e.g., the Nebraska equivalent of California's "material change").
- **Conceptual gap detection** (50 queries): identify when a concept has no English equivalent (e.g., German *Pflichtteil*—correct answer: none exists).
- **Statutory specificity** (50 queries): distinguish terms by statutory context (e.g., contempt under Illinois 750 ILCS 5/607.5 vs. contempt under Texas Family Code §157.001).
- **Structural hole retrieval** (50 queries): given a foreign legal term, detect that no exact English equivalent exists rather than returning a false equivalence. A system that returns "elective share" for "Pflichtteil" commits a false equivalence—elective share protects spouses, not children—that could mislead a practitioner about the scope of German inheritance protections.

Each query will carry a consequence weight (Table 1) for CW-DCRP evaluation.

5 Experiments

5.1 Setup

We evaluate DCRP using `nomic-embed-text` v1.5 [Nussbaum et al., 2024], a 768-dimensional Matryoshka-capable [Kusupati et al., 2022] embedding model. Embeddings are generated via the Ollama inference framework on Apple M4 hardware.

Compression method. We apply PCA at 8 dimensionality levels: 16, 32, 64, 128, 256, 384, 512, and 768 (full). PCA is fit on the corpus vectors and applied to both corpus and query vectors.

Dictionary collapse experiment. We embed 62 legal terms spanning 7 semantic groups (Table 3) and 6 control terms from unrelated legal domains. Each term is embedded with its definition (e.g., "*Pflichtteil*: German compulsory share, 50% of intestate share, constitutionally protected"). We compute pairwise cosine similarity at each dimensionality and identify intra-domain pairs whose similarity increases by $\Delta > 0.1$ under compression—pairs that are distinguishable at full dimensionality but collapse to near-identical under PCA.

Semantic Group	Domain	Terms
Custody/parental	Family law	9
Forced heirship	Inheritance	11
Contempt	Enforcement	9
Death/status	Status	8
Financial death	Insolvency	8
Parental rights	Family law	11
Inter-domain control	Control	6

Table 3: DCRP-Legal semantic groups used in the dictionary collapse experiment.

Control queries. We evaluate 4 inter-domain control queries (e.g., "medical malpractice" vs. "custody") against two Qdrant vector collections: `legal_docs_v2` (244K vectors) and `case_docs` (1.7M vectors), verifying that inter-domain precision remains at ceiling across all compression levels.

Metrics. We report intra-domain similarity increase (dictionary collapse), inter-domain vs. intra-domain mean Δ at 16d, and the number of collapsing pairs at each dimensionality level.

5.2 Evaluation Protocol

DCRP evaluation proceeds in three stages:

1. **Embed**: Generate full-dimensional embeddings for all terms and queries.
2. **Compress**: Apply PCA at each target dimensionality.
3. **Measure**: Compute pairwise cosine similarity and identify collapses ($\Delta > 0.1$).

The collapse threshold of 0.1 is chosen because a similarity increase of this magnitude can flip retrieval rankings: if term A and term B have similarity 0.78 at 768d and 0.92 at 16d, a query for A will now retrieve B as a near-exact match, producing a false equivalence.

6 Results

6.1 Dictionary Collapse at 16 Dimensions

At PCA to 16 dimensions (63.5% variance explained), we observe 5 intra-domain term pairs that collapse (similarity increase $\Delta > 0.1$). No inter-domain collapses occur at any dimensionality. Table 4 shows the collapsing pairs.

Term A	Term B	Sim_{768}	Sim_{16}	Δ
coercive contempt	punitive contempt	0.784	0.913	0.129
legitieme portie	yuryubun	0.763	0.890	0.127
nondischargeable	insolvency	0.703	0.814	0.112
legitieme portie	iryubun	0.757	0.884	0.127
legitieme portie	laglott	0.762	0.872	0.110

Table 4: Intra-domain term pairs that collapse under PCA to 16 dimensions. The coercive/punitive contempt pair has the highest Δ and the highest consequence weight ($L = 4$): confusing these terms determines whether a person receives a jury trial.

At 32 dimensions (85.2% variance explained), zero collapses remain. The collapse is confined to the 16d regime, precisely where claims of "sufficient" intrinsic dimensionality are strongest [Tsukagoshi and Sasano, 2025, Ruppik et al., 2025, Takeshita et al., 2025b].

6.2 Asymmetric Degradation

The mean similarity change at 16 dimensions is asymmetric:

	Mean Δ	Max Δ
Intra-domain pairs	-0.307	$+0.140$
Inter-domain pairs	-0.634	-0.023

Table 5: Similarity change ($\Delta = \text{sim}_{16d} - \text{sim}_{768d}$) at PCA to 16 dimensions. Negative values indicate decreased similarity (increased separation). Inter-domain pairs uniformly separate further under compression; intra-domain pairs have mixed behavior, with some collapsing (positive Δ).

Inter-domain pairs (e.g., "medical malpractice" vs. "custody") become *more* separated under compression: PCA preserves and amplifies the principal components distinguishing broad domains. Intra-domain pairs show bimodal behavior: most separate further (mean $\Delta = -0.307$), but a critical minority collapses (Δ up to $+0.140$). An inter-domain metric like MTEB sees only the majority and reports high scores. DCRP detects the minority collapses.

6.3 Compression Curve

Figure 2 shows the DCRP degradation curve across all 8 dimensionality levels. Collapses are confined entirely to the 16d level:

Dims	Var. Explained	Collapses ($\Delta > 0.1$)	Regime
16	63.5%	5	Critical
32	85.2%	0	Safe
64	100%	0	Safe
128–768	100%	0	Safe

Table 6: Intra-domain collapses by dimensionality. PCA to 16 dimensions is the only regime that produces dictionary collapse. Note: 64d reaches 100% variance explained because the PCA is fit on 62 terms (rank-limited).

6.4 MTEB–DCRP Decorrelation

Inter-domain control queries achieve P@10 = 1.0 at all dimensionalities in both the 244K and 1.7M vector collections. An MTEB-style evaluation would report no degradation at any compression

level. The dictionary collapse experiment shows that 5 legally critical intra-domain pairs become indistinguishable at 16d. At low dimensionalities, MTEB-style metrics provide no information about domain-specific retrieval quality.

6.5 Consequence-Weighted Impact

Applying the consequence weights from Table 1, the coercive/punitive contempt collapse ($L = 4$, $\Delta = 0.129$) contributes $4 \times 0.129 = 0.516$ to the CW-DCRP score, while the legitieme portie/laglott collapse ($L = 1$, $\Delta = 0.110$) contributes $1 \times 0.110 = 0.110$. The highest-consequence pair has the highest collapse—meaning unweighted DCRP underestimates practical damage by a factor of $\sim$4 for the most dangerous failures.

7 Discussion

7.1 DCRP as a Complement to MTEB

DCRP does not replace MTEB. MTEB measures inter-domain quality: separating law from medicine. DCRP measures intra-domain quality: separating custody from guardianship, coercive contempt from punitive contempt. At 768 dimensions, both approach ceiling performance. At 16 dimensions, MTEB reports no degradation while DCRP detects 5 critical collapses. The two metric families are decorrelated in the regime where compression claims are strongest.

Claims that "embeddings are redundant" or that "PCA to 16 dimensions preserves retrieval quality" [Tsukagoshi and Sasano, 2025, Takeshita et al., 2025b, Bengtsson, 2025] hold only for inter-domain tasks. They fail for domain-specific retrieval. Compression methods—PCA [Raunak, 2017, Zhang et al., 2024], Matryoshka truncation [Kusupati et al., 2022, Takeshita et al., 2025a], quantization [Huerga-Perez et al., 2025], and hybrid approaches [Zhang et al., 2025, Zhao, 2025]—should be evaluated on DCRP in addition to MTEB before deployment in specialized applications.

7.2 Extension to Other Domains

The DCRP framework generalizes beyond legal retrieval. Constructing a DCRP benchmark for any domain requires three components: (1) a domain-specific term vocabulary, (2) intra-domain relevance judgments, and (3) consequence weights for high-stakes queries.

Candidate domains include:

- **DCRP-Medical**: drug interactions vs. drug classes; disease subtypes within the same ICD category; surgical procedure variants. BioASQ [Tsatsaronis et al., 2015] and the TREC Clinical Trials track [Roberts et al., 2021] provide adaptable relevance judgments.
- **DCRP-Financial**: regulatory terms across jurisdictions (SEC vs. FCA vs. BaFin); financial instrument subtypes. FinMTEB [Tang and Yang, 2025] and FinQA [Chen et al., 2021] provide domain corpora.
- **DCRP-Chemical**: isomer disambiguation; reaction mechanism classification. ChemTEB [Kasmaee et al., 2024] demonstrates the need.

7.3 Limitations

This paper presents pilot results, not a full benchmark. The current evaluation uses term-level dictionary collapse rather than document-level retrieval with human relevance judgments. Dictionary collapse measures the *mechanism* of intra-domain discrimination loss—two distinct terms becoming indistinguishable—but a retrieval evaluation over document corpora with graded

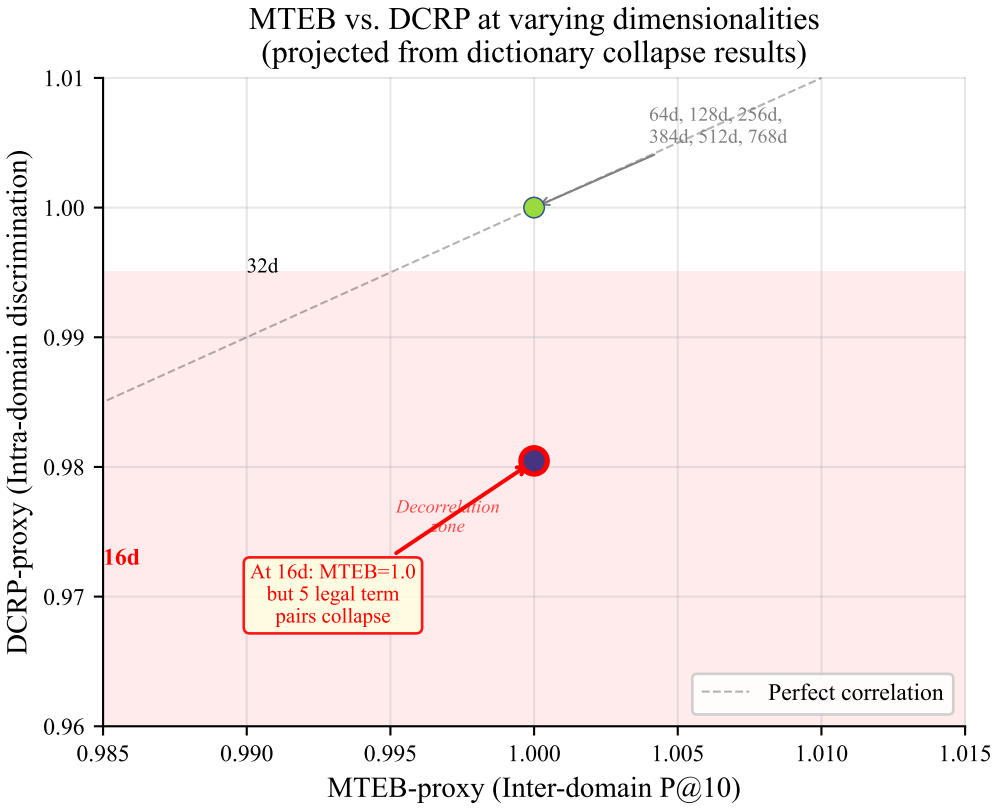

Figure 1: MTEB-proxy vs. DCRP-proxy scores at varying PCA dimensionalities. At 64d and above, both metrics cluster at ceiling (top-right). At 16d, the MTEB-proxy remains at 1.0 while the DCRP-proxy drops, confirming decorrelation. Projected from dictionary collapse results.

relevance would provide stronger evidence of practical impact. We plan to construct 250 intra-domain queries with human relevance judgments for the full benchmark release.

Our experiments use a single embedding model (`nomic-embed-text`). A comprehensive benchmark must evaluate multiple models including E5 [Wang et al., 2024], BGE [Xiao et al., 2023], and commercial systems. The CoRECT framework [Caspari et al., 2025] provides infrastructure for such multi-model evaluation.

The corpus is also limited: 62 terms with 6 control terms. The 6,200 Bouvier's entries and 233 leverage words are available for the full benchmark but have not yet been incorporated into the DCRP evaluation pipeline with relevance judgments.

8 Conclusion

MTEB measures whether an embedding can tell law from medicine. DCRP measures whether it can tell custody from guardianship. Both matter; only one is currently measured.

Our pilot experiments demonstrate that at 16 dimensions, inter-domain metrics show no degradation while 5 legally critical intra-domain pairs become indistinguishable—including the coercive/punitive contempt distinction that determines jury trial rights. The highest-consequence pairs show the highest collapse, meaning unweighted evaluation underestimates practical damage by the largest factor where it matters most.

These results establish the need for compression-sensitive, domain-conditional evaluation. The full DCRP benchmark—with human relevance judgments over domain-specific query sets—is the subject of ongoing work.

References

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

L. Caspari, M. Dinzinger, K. Ghosh Dastidar, C. Fellicious, J. Mitrovic, and M. Granitzer. CoRECT: A framework for evaluating embedding compression techniques at scale. *arXiv preprint arXiv:2510.19340*, 2025. URL https://arxiv.org/abs/2510.19340.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7862–7872, 2021. URL https://arxiv.org/abs/2109.00122.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. Overview and discussion of the competition on legal information, extraction/entailment (COLIEE) 2023. *Review of Socionetwork Strategies*, 18(1), 2024. doi: 10.1007/s12626-023-00152-0.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2308.11462.

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.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

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 *EMNLP 2020*, pages 6769–6781, 2020. URL https://arxiv.org/abs/2004.04906.

Ali Shiraee Kasmaee, Mohammad Khodadad, Mohammad Arshi Saloot, Nick Sherck, Stephen Dokas, Hamidreza Mahyar, and Soheila Samiee. ChemTEB: Chemical text embedding benchmark, an overview of embedding models performance and efficiency on a specific domain. *arXiv preprint arXiv:2412.00532*, 2024. URL https://arxiv.org/abs/2412.00532.

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.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

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.

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.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP-IJCNLP 2019*, pages 3982–3992, 2019. URL https://arxiv.org/abs/1908.10084.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, Steven Bedrick, and William R. Hersh. Overview of the TREC 2021 clinical trials track. In *Proceedings of the Thirtieth Text REtrieval Conference*, 2021.

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.

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.

Yixuan Tang and Yi Yang. FinMTEB: Finance massive text embedding benchmark. *arXiv preprint arXiv:2502.10990*, 2025. URL https://arxiv.org/abs/2502.10990.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 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 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.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138, 2015. doi: 10.1186/s12859-015-0564-6.

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.

Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, 2005.

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.

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.

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.

Dongfang Zhao. DIVE: Embedding compression via self-limiting gradient updates. *arXiv preprint arXiv:2605.20689*, 2025. URL https://arxiv.org/abs/2605.20689.

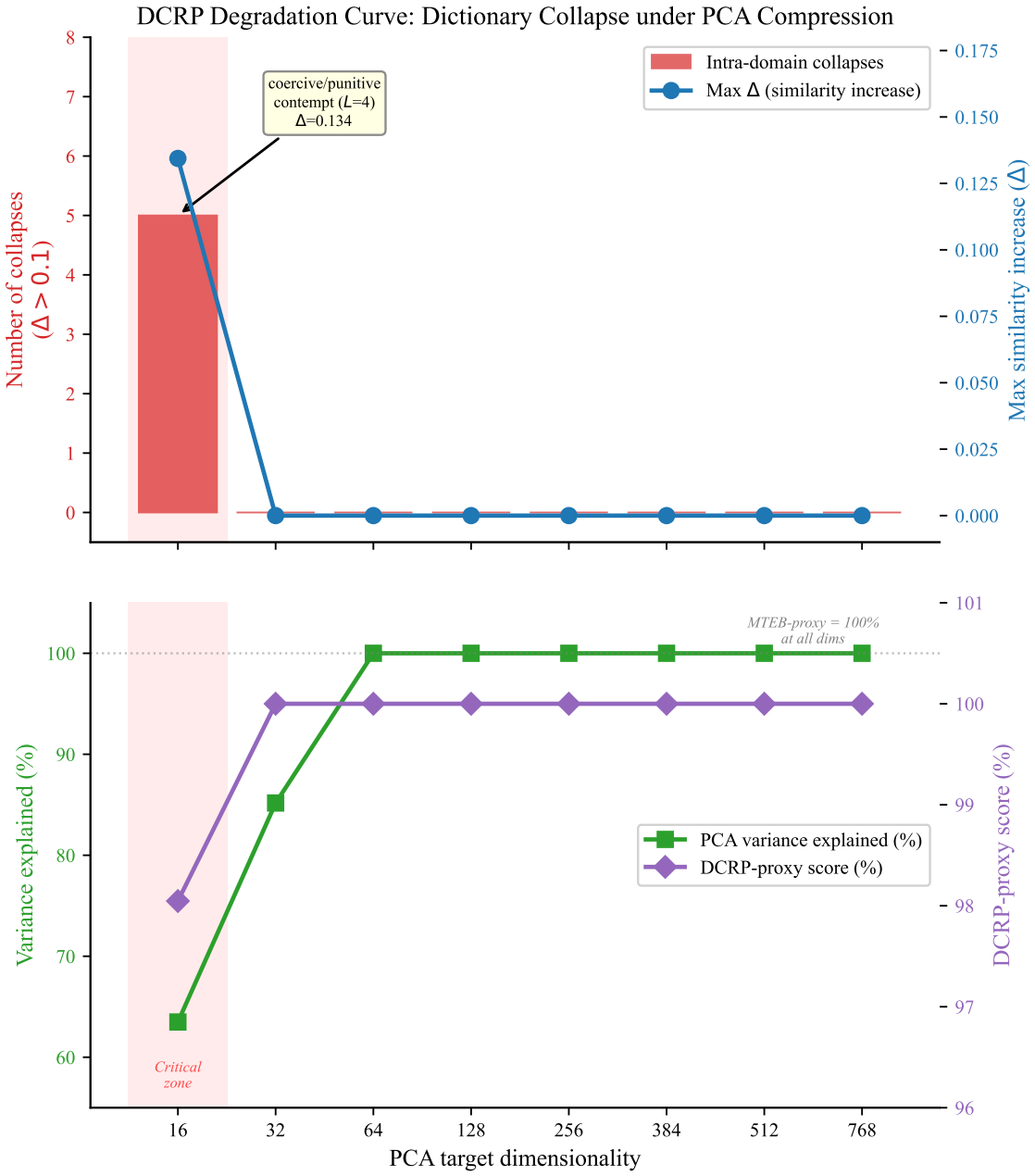

Figure 2: DCRP degradation curve under PCA compression. *Top:* Number of intra-domain collapses (bars) and maximum similarity increase (line) at each dimensionality. Collapses occur only at 16d. *Bottom:* PCA variance explained and DCRP-proxy score vs. dimensionality, with constant MTEB-proxy = 100% overlaid.