

Beyond Compressibility: Transferable Low-Dimensional Structure and the Limits of Compression in Financial Retrieval

Oddharak Tyagi, Shashishekar Ramakrishna, Chetan Harsha,
Sridhar Dasaratha

EY Global Delivery Services India LLP

{Oddharak.Tyagi, Shashishekar.R, Chetan.Harsha,
Sridhar.Dasaratha}@gds.ey.com

Abstract

Modern retrieval embeddings are highly compressible, often retaining most of their effectiveness after the removal of a large fraction of dimensions. This suggests that retrieval-relevant information is organized in a lower-dimensional structure, but it remains unclear how broadly that structure generalizes and what ultimately limits compression.

We study this question in long-document financial retrieval where relevant evidence must be distinguished from many highly similar passages containing repeated entities, financial metrics, tables, and boilerplate. Across three financial retrieval benchmarks and eight retrievers, we find that PCA projections learned on one corpus transfer effectively to others, retaining most of the effectiveness of corpus-specific fitting. Transfer quality is strongly associated with principal-subspace overlap, consistent with substantial shared low-dimensional structure across financial corpora. We further show that global geometric measures such as intrinsic dimension only partially explain retrieval robustness under dimensionality reduction. A case study on a financial question answering benchmark links retrieval failures to dense clusters of competing table chunks: Removing these distractors substantially reduces performance degradation. These results suggest that transferability is enabled by low-dimensional structure that generalizes across financial corpora, while compression failures are associated with local retrieval competition that is not captured by existing global geometric measures.

1 Introduction

Dense retrieval has become a central component of modern information retrieval and retrieval-augmented generation (RAG) systems (Karpukhin et al., 2020; Lewis et al., 2020). The success of this paradigm has led to its adoption in domains where retrieving evidence from large collections of

unstructured documents is required. For instance, financial analysts routinely search lengthy annual reports, earnings filings and regulatory documents to locate supporting evidence.

The financial QA capabilities of LLMs over short curated excerpts have been well studied (Phogat et al., 2023), and recent benchmarks (Islam et al., 2023; Wang et al., 2026; Moslonka et al., 2026) have shifted evaluation towards retrieval and QA over complete financial documents, often hundreds of pages long (>100k tokens). Financial reports contain extensive boilerplate, repeated entities, recurring numerical disclosures, and large amounts of tabular content which create many highly similar candidate passages. This is why passage retrieval from long financial documents remains a challenging problem (Kim et al., 2025; Kobeissi and Langlais, 2026).

Recent studies have shown that text embeddings contain substantial redundancy, allowing dimensionality reduction with little loss in retrieval effectiveness (Takeshita et al., 2025), and in some cases even some domain adaptation (Zuo and Khashabi, 2026). These findings suggest that retrieval-relevant information may occupy a much lower-dimensional space than the embeddings themselves. Principal Components Analysis (PCA) (Abdi and Williams, 2010) has emerged as a simple and effective tool for studying this phenomenon, but prior work fits the projection directly on the evaluation corpus. As a result, it remains unclear whether successful compression reflects corpus-specific artifacts or a more general reusable low-dimensional structure. Compression therefore provides a useful lens for studying the geometry of retrieval representations rather than merely a tool for reducing storage and computation.

If retrieval-preserving compression reflects genuine low-dimensional structure, then that structure should generalize across related retrieval corpora. We therefore ask whether low-dimensional projec-

tions learned on one financial corpus remain effective on another. This question is practically important because retrieval systems often operate over evolving document collections where repeated re-fitting may be undesirable. More fundamentally, successful transfer would indicate that compression is exploiting structure that generalizes across corpora rather than artifacts specific to the dataset used for fitting.

A second question concerns when dimensionality reduction fails. Recent work has linked representation redundancy to global geometric properties such as intrinsic dimension (Tsukagoshi and Sasano, 2025). However, retrieval quality depends on preserving distinctions between relevant evidence and competing candidates, raising the possibility that retrieval robustness is governed by more than corpus-level geometry.

In this work, we investigate the structure underlying retrieval-preserving compression. Our primary question is whether PCA projections learned on one financial corpus can be transferred to another, and whether transferability is related to the alignment of their principal subspaces. We then use compression failures to probe the limits of this shared structure, asking whether retrieval robustness is explained by global geometric properties or by the local structure of competing retrieval candidates.

Our contributions are as follows:

- We provide a systematic empirical study of PCA-based dimensionality reduction for long financial document retrieval across three benchmarks and eight retrievers. We find that retrieval performance is largely preserved, and sometimes even improved, under substantial dimensionality reduction.
- We show that PCA projections learned on one financial corpus transfer effectively to other financial corpora. Transfer performance is strongly associated with principal-subspace overlap, suggesting that these datasets share low-dimensional structure that can be exploited without fitting the compression transform directly on the target corpus.
- We show that measures such as explained variance and intrinsic dimension only partially explain retrieval compressibility. Although these measures capture broad trends, they do not fully predict robustness within long-document financial retrieval.

- We identify a structured compression failure mode in DocFinQA (Reddy et al., 2024): aggressive dimensionality reduction disproportionately harms retrieval when relevant evidence must be distinguished from many highly similar competing table chunks, highlighting the role of local retrieval competition in compression robustness.

2 Related Work

Embedding compression for retrieval. A substantial literature studies how to reduce the storage and computational cost of dense retrieval. Existing approaches include quantization, autoencoders, and representation learning methods that explicitly support variable embedding sizes (Zouhar et al., 2022; Liu et al., 2022; Kusupati et al., 2022). PCA has emerged as a simple and effective post-hoc dimensionality reduction technique for retrieval, preserving effectiveness under substantial compression and recently being explored for efficient financial RAG systems (Ma et al., 2021; Khaledian et al., 2025).

Most prior work evaluates dimensionality reduction on the same corpus used to estimate the projection matrix, making it difficult to determine whether successful compression reflects corpus-specific structure or reusable low-dimensional organization. In contrast, we explicitly study cross-corpus transfer of PCA projections and ask whether compression transforms learned on one corpus remain effective on another.

Redundancy in text embeddings. Recent work has shown that modern embedding models contain substantial redundancy. Takeshita et al. (2025) find that removing a large fraction of embedding dimensions often has surprisingly little impact on retrieval and classification performance. In dense retrieval, Zuo and Khashabi (2026) demonstrate that PCA compression can preserve and sometimes improve effectiveness, interpreting dimensionality reduction as a lightweight form of domain adaptation. Their approach fits PCA using target-domain query embeddings, whereas we focus on document-side fitting and evaluate whether projections learned from one corpus transfer to entirely new collections without access to target-domain queries.

Most prior studies focus on general-domain retrieval benchmarks (Thakur et al., 2021; Muenighoff et al., 2023), whereas we examine compression in long-document financial retrieval con-

taining extensive boilerplate, repeated entities, and tabular content.

Embedding geometry and post-processing.

Text embeddings are known to be highly anisotropic, motivating post-processing methods such as mean removal, common-component removal, whitening, and isotropy correction (Ethayarajh, 2019; Mu and Viswanath, 2018; Su et al., 2021; Huang et al., 2021; Li et al., 2026). However, the effectiveness of these methods depends strongly on the embedding model and downstream task (Forooghi et al., 2024; Ait-Saada and Nadif, 2023).

More recently, studies have connected embedding compressibility to geometric properties such as isotropy and intrinsic dimension. Tsukagoshi and Sasano (2025) find that lower intrinsic dimension is associated with greater compressibility across embedding tasks. However, existing work largely studies representation-level properties rather than retrieval behavior itself. In this paper, we ask whether global geometric measures are sufficient to explain retrieval robustness under compression, and find that local competition among highly similar retrieval candidates appears more closely associated with compression failures in long-document financial retrieval.

3 Experiments

3.1 Datasets

We evaluate on three finance-domain retrieval benchmarks: FiQA (Maia et al., 2018), DocFinQA (Reddy et al., 2024), and ESG-Bench (Sun et al., 2026).

FiQA is a short-passage retrieval benchmark derived from financial forums. **DocFinQA** evaluates evidence retrieval from long financial filings and annual reports. **ESG-Bench** focuses on retrieval from long-form environmental, social, and governance reports. For DocFinQA and ESG-Bench, retrieval is performed over chunked document representations, with relevant chunks/pages defined by the released evidence annotations for each benchmark.

For computational efficiency, we randomly sample filings from the official DocFinQA test split. Chunk-level retrieval labels are constructed using an evidence-alignment procedure and filtered using a confidence threshold, resulting in 417 retained question-answer retrieval instances (Appendix B). ESG-Bench contains no official train/test split; we evaluate on 251 reports (251 queries) and create a

document-level 80/20 split solely for in-distribution PCA experiments. For FiQA, we follow the standard evaluation protocol and use the complete benchmark without modification. Dataset statistics are provided in Appendix A.

3.2 Retriever Models

We evaluate eight dense retrievers spanning a range of model sizes, embedding dimensions, and training objectives. Embeddings are generated using each model’s recommended retrieval prompts and L2-normalized before PCA fitting and retrieval. Full model details are provided in Table 5.

3.3 PCA Fitting Protocol

To study how redundancy depends on the data used to estimate the projection matrix, we study PCA fitting in three settings: **In-sample:** PCA is fit and evaluated on the same corpus. **In-distribution:** PCA is fit on a held-out training corpus from the same dataset and applied unchanged at evaluation. **Out-of-distribution:** PCA is fit on one dataset and transferred directly to another (for instance, DocFinQA $\rightarrow$ ESG-Bench). Queries are not used for fitting, since target-domain queries may not be available in many real-world retrieval deployments. PCA is therefore estimated solely from corpus embeddings.

The out-of-distribution setting provides a direct test of whether effective compression relies primarily on corpus-specific regularities or on structure that generalizes across financial retrieval datasets.

For FiQA, PCA is fit on document embeddings. For DocFinQA and ESG-Bench, PCA is fit on 100k sentence-level chunk embeddings sampled uniformly from the fitting corpus and produced using a 150 max-token sentence chunker. The learned projection matrix is applied identically to query and document embeddings prior to retrieval. Additional implementation details and fitting-granularity ablations are provided in Appendix D.

3.4 Metrics

We report nDCG@10 as the primary retrieval metric and provide Recall@10 results in Appendix I. In-distribution results for DocFinQA are provided unless otherwise specified. Other datasets evaluated in-sample. Statistical significance is assessed using paired bootstrap resampling over query-level differences with 10k samples. We report 95% confidence intervals and consider differences significant when the interval excludes zero.

4 Results

4.1 Retrieval embeddings exhibit substantial redundancy

Figure 1 shows that retrieval performance is largely preserved under substantial PCA-based dimensionality reduction across all three benchmarks. Even after retaining only 10% of the original embedding dimensions, average nDCG@10 decreases by roughly 10%. Averaging across retrievers understates the best case: for individual retrievers, compression frequently *improves* retrieval over the uncompressed baseline. On ESG-Bench, 7 of the 8 retrievers exceed their full-dimensional nDCG@10 at some compression budget, and Granite-s-r2 gains as much as 7.7% at $k/D = 0.15$ (an 85% reduction in dimensionality). Even on the harder DocFinQA benchmark, 2 retrievers surpass their baseline under compression, with Granite-r2 improving by up to 5.0% at $k/D = 0.3$. These best-case gains show that PCA can not only preserve but occasionally improve retrieval effectiveness. DocFinQA and FiQA display surprisingly similar degradation curves despite representing very different retrieval settings. Results for individual retrievers along with baselines can be found in Appendix I.

Appendix C shows that PCA fitted on a disjoint in-distribution corpus achieves nearly identical performance to PCA fitted directly on the evaluation corpus. This result suggests that the learned projection is not over-specialized to the documents used during fitting. From a practical perspective, it indicates that a projection estimated once can be reused on newly arriving filings or reports from the same distribution, reducing the need for repeated refitting. Prior work has largely evaluated PCA-based dimensionality reduction in the in-sample setting, making this a useful prerequisite for the cross-dataset transfer experiments that follow.

4.2 Transferability across financial corpora

The dimensionality reduction results above suggest that retrieval-preserving information occupies a low-dimensional subspace. A natural question is whether this subspace is specific to each corpus or whether it reflects structure shared across financial retrieval tasks. To test this hypothesis, we fit PCA on a source dataset and apply the found projection to a different target corpus. Figure 2 reports, for each directed pair of corpora, the change in nDCG@10 of this transferred projection relative to the target’s own native PCA as a function of the

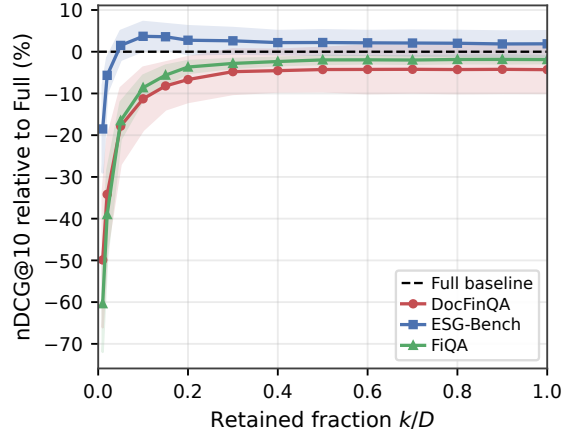


Figure 1: Relative change in nDCG@10 with respect to full-dimensional baseline. Performance averaged across retrievers with ± 1 standard deviation.

retained dimension, so that a small gap indicates that the low-dimensional structure is shared rather than corpus-specific.

Figure 2 shows that transferred projections perform remarkably close to native PCA projections. Even at 80% compression, the largest reduction in nDCG@10 relative to native projection is only around 3%. Transfer remains effective even between FiQA and the long-document datasets despite their substantial differences.

The gap between native and transferred PCA widens under extreme compression, but transferred projections consistently outperform both random projections and Matryoshka truncation across all compression budgets. The non-zero change in performance even at full dimension ($k/D = 1$) arises from the mean centering used in PCA.

These results indicate that a substantial portion of the retrieval-preserving structure captured by PCA is not corpus-specific. Instead, the dominant low-dimensional directions learned from one financial corpus remain useful when applied to another.

What determines transferability? A natural explanation for the strong transfer results is that different financial corpora occupy similar principal subspaces. Under this hypothesis, transfer succeeds because the dominant directions retained by PCA largely coincide across corpora. A projection learned on one dataset therefore removes approximately the same variation as a projection learned on another.

To quantify this similarity, we measure subspace overlap using the mean squared cosine of the prin-

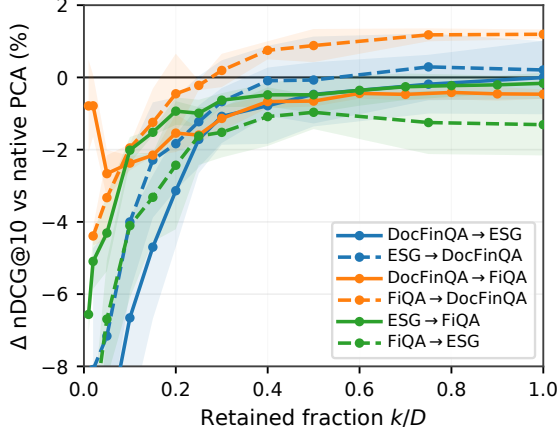


Figure 2: Change in nDCG@10 between transferred and natively fitted PCA projection. The performance of transferred PCA projections is $\lesssim 3\%$ of native PCA until $k/D \sim 0.2$, indicating that dominant variance directions are significantly shared across financial datasets.

principal angles between the top- k PCA subspaces:

$$\overline{\cos^2}(A, B; k) = \frac{1}{k} \left\| U_A^{(k)\top} U_B^{(k)} \right\|_F^2.$$

Higher values indicate greater overlap between the dominant embedding directions of two corpora.

If transfer succeeds because two corpora share similar dominant directions, then corpora with more aligned principal subspaces should exhibit smaller transfer losses. Figure 3 shows that across financial corpora, subspace overlap is strongly correlated with transfer performance at $k/D = 0.1$ (Pearson $r = +0.82$, Spearman $\rho = +0.63$; $n = 24$ transfer settings, as FiQA is embedded with only two of the eight retrievers): higher overlap is consistently associated with smaller transfer losses. The same relationship holds at other k/D . These points are not independent, as each of the 8 retrievers contributes several transfer directions; a cluster bootstrap over retrievers (Pearson $[+0.50, +0.93]$) and leave-one-retriever-out (Pearson $[+0.70, +0.86]$) confirm the association is not driven by any single retriever (Appendix E).

The transfer experiments suggest that PCA compression is not merely exploiting corpus-specific regularities. Low-dimensional retrieval subspaces learned on one financial corpus remain highly effective on others, and transfer performance is associated with subspace alignment.

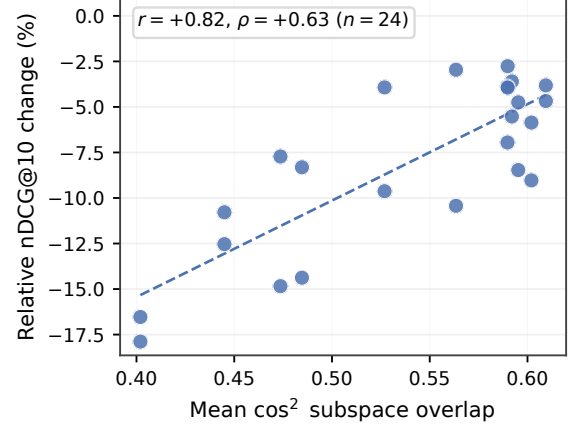


Figure 3: Source–target principal-subspace overlap versus transfer degradation at $k/D = 0.1$. Each point is a retriever–transfer setting. Higher mean- $\cos^2$ overlap is associated with smaller transfer loss.

4.3 Intrinsic dimension is insufficient to explain retrieval compressibility

A natural explanation for the strong compression results is that some datasets simply occupy lower-dimensional regions of the embedding space than others. Recent work has linked embedding compressibility to geometric properties such as intrinsic dimension (Tsukagoshi and Sasano, 2025), suggesting that datasets with lower effective dimensionality should tolerate more aggressive compression. We therefore ask whether the compression behavior observed in financial retrieval can be explained by standard geometric descriptors of the embedding distribution.

We begin with the covariance eigenspectrum, which provides a simple measure of how concentrated the embedding distribution is. If retrieval compressibility were primarily determined by global redundancy, datasets whose variance is concentrated in fewer principal directions should tolerate compression more gracefully. Figure 4 shows that DocFinQA and ESG-Bench exhibit much faster spectral decay than FiQA. However, this ordering does not mirror retrieval robustness, suggesting that explained variance alone is insufficient. Notably, the variance-explained curves are remarkably similar within a dataset despite large differences in embedding dimensionality and training objectives (Appendix G).

Intrinsic dimension (ID) provides a richer geometric characterization than explained variance. We estimate ID using the TwoNN estimator (Facco et al., 2017) and compare PCA, MRL truncation,

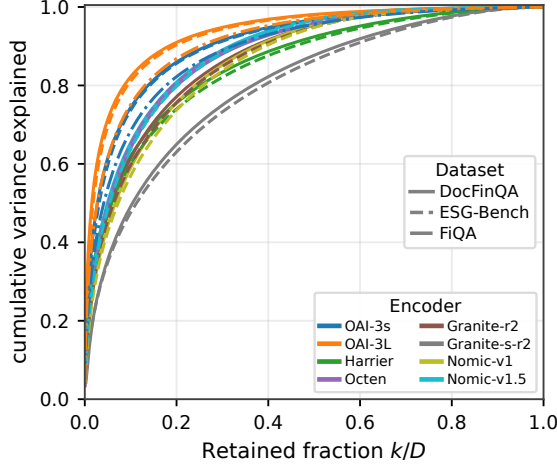


Figure 4: Cumulative variance explained as a function of retained dimensions k/D . DocFinQA and ESG-Bench exhibit faster spectrum decay than FiQA.

and random projection in Figure 5. The datasets exhibit clear geometric differences: DocFinQA has the lowest estimated ID, ESG-Bench is intermediate, and FiQA the highest. However, this ordering again fails to predict retrieval compressibility. ESG-Bench is the most compressible dataset despite not having the lowest ID, while DocFinQA and FiQA display similar compression behavior despite substantially different estimated dimensions. Interestingly, PCA substantially lowers ID at matched dimension budgets, whereas MRL truncation and random projection largely preserve it. Global geometry therefore captures representation redundancy, but is insufficient to explain retrieval robustness under dimensionality reduction.

One possible explanation is that retrieval depends on local rather than corpus-level geometry. We therefore estimate ID within query-centered neighborhoods using the maximum-likelihood estimator of [Levina and Bickel \(2004\)](#) implemented in `scikit-dimension` ([Bac et al., 2021](#)). Local dimensions are considerably lower than corpus-level estimates, but they do not resolve the discrepancy. DocFinQA and ESG-Bench have similar local dimensions despite substantially different compression behavior (Appendix H), and across all 18 retriever–dataset pairs (FiQA is embedded with only two of the eight retrievers) neither global nor local ID is meaningfully correlated with the compression knee¹ (Pearson -0.05 , Spearman -0.24).

¹The compression knee is the smallest retained fraction k/D at which retrieval quality remains within 95% of the uncompressed (full-dimensional) baseline.

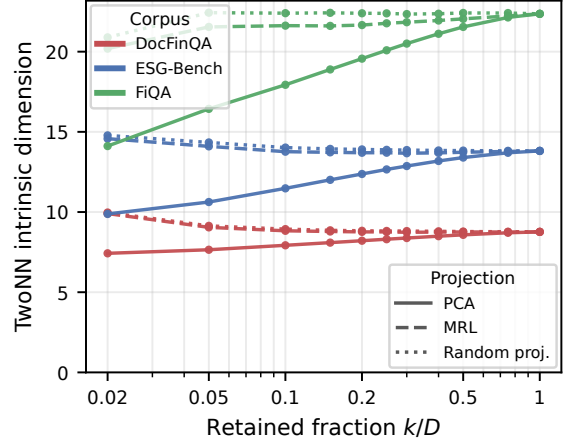


Figure 5: Intrinsic Dimension (ID) as a function of retained fraction k/D under PCA, MRL truncation, and random orthonormal projection. At matched budget PCA lowers the ID, whereas MRL and random projection largely preserve it.

The failure of both global and local ID suggests that dimensionality alone may not be the relevant quantity. Retrieval quality depends on separating relevant evidence from competing candidates, suggesting that the density of local neighborhoods may matter more than their intrinsic dimensionality. As a simple proxy for local retrieval competition, we compute the average cosine similarity between each query and its nearest neighbors. Figure 6 reveals a pattern that more closely matches retrieval behavior. Dimensionality reduction also increases the similarity of nearby candidates, reducing the separation between relevant passages and distractors.

Unlike classification, retrieval is inherently relational: success depends on preserving relative distances among competing candidates rather than simply maintaining a separable representation space. This distinction may explain why intrinsic dimension captures important aspects of representation compressibility while remaining weakly associated with retrieval compressibility. In retrieval, even a highly compressible embedding distribution can produce substantial performance degradation if dimensionality reduction erodes the ranking margins that separate relevant evidence from distractors. The next section examines this possibility directly by analysing which competing candidates are promoted when compression fails.

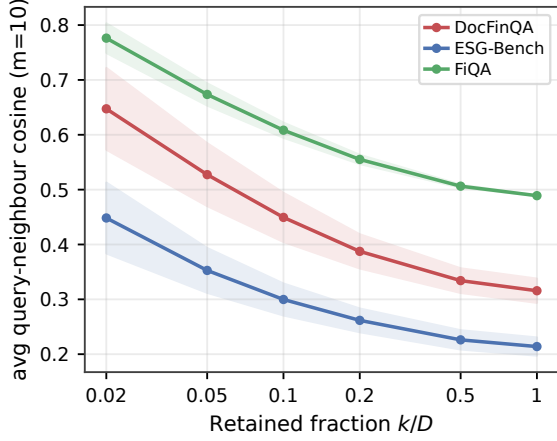


Figure 6: Mean cosine of a query to its 10 nearest neighbors as a function of retained fraction k/D , averaged over retrievers and queries (band: ± 1 standard deviation across retrievers).

4.4 What drives retrieval compressibility?

DocFinQA differs from ESG-Bench not only in its global geometry but also in the structure of its candidate pool. Financial filings frequently interleave narrative text with large serialized tables, creating many highly similar retrieval candidates. To examine whether local candidate structure contributes to DocFinQA’s reduced compressibility, we perform a focused case study using `text-embedding-3-small`.

We label a chunk as *table* when it contains the chunker’s serialized column markers and *no-table* otherwise, yielding a table rate of 21% across DocFinQA. Gold retrieval chunks are aligned to the original FinQA evidence spans using lexical overlap. Only about one-third of aligned gold chunks (131 of 417) contain a table, while others correspond to narrative explanations or discussions surrounding financial figures. To isolate table-specific effects, all query-level analyses below are restricted to the 131 queries whose aligned gold evidence is tabular.

Table chunks form dense local neighborhoods.

Table chunks are substantially more similar to one another than no-table chunks. At $k/D = 0.02$, the mean pairwise cosine similarity is 0.36 for tables and 0.18 for no-table chunks (0.31 versus 0.15 at full dimension). Tables are also strongly self-clustered: among a table chunk’s top-5 nearest neighbors within the same document, between 0.55 and 0.61 are themselves tables, far above the corpus base rate of 0.21. Retrieval therefore often requires

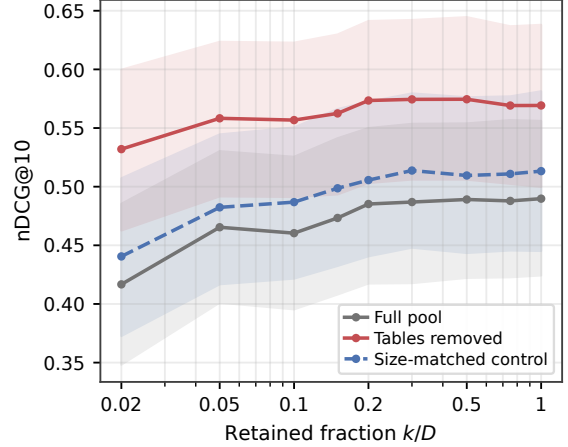


Figure 7: Removing competing table chunks reduces the performance loss caused by dimensionality reduction in DocFinQA. The size-matched control removes an equal number of random no-table chunks (mean of 5 seeds). Bands denote 95% bootstrap confidence intervals.

distinguishing a relevant table from many highly similar alternatives.

Table competition amplifies compression failures.

We test whether these dense neighborhoods contribute to compression failures by removing all *non-gold* table chunks from the candidate pool while retaining the gold chunk. As a control, we remove an equal number of randomly selected no-table chunks. The size-matched control distinguishes the effect of removing table competitors from the effect of simply reducing the candidate pool.

Figure 7 shows a substantial effect. At $k/D = 0.02$, removing competing table chunks increases $nDCG@10$ from 0.42 to 0.53, compared with 0.44 for the size-matched control. Compression sensitivity also decreases: the drop in $nDCG@10$ between full dimension and $k/D = 0.02$ falls from 0.073 to 0.037.

These findings suggest that retrieval compressibility depends not only on the global geometry of the embedding distribution but also on the density of competing candidates. In DocFinQA, dimensionality reduction is most harmful when relevant evidence must be separated from dense clusters of highly similar table chunks, a factor that is not captured by corpus-level measures such as intrinsic dimension.

5 Discussion

A central finding of this work is that representation compressibility and retrieval compressibility are related but distinct phenomena. Financial retrieval corpora exhibit substantial shared low-dimensional structure: PCA projections learned on one corpus transfer effectively to others, and transfer performance is strongly associated with principal-subspace overlap. This suggests that much of the retrieval-preserving information captured by PCA lies in low-dimensional directions that generalize across financial corpora.

However, preserving this global structure is not sufficient to preserve retrieval quality. Although measures such as explained variance and intrinsic dimension capture broad properties of the embedding distribution, they do not fully explain how retrieval degrades under compression. In particular, DocFinQA and ESG-Bench exhibit markedly different compression robustness despite showing similar behavior under global geometric measures.

Our case study points to a complementary explanation. Compression can preserve much of a corpus’s global geometry while still degrading the local ranking margins needed to separate relevant evidence from highly similar alternatives. In DocFinQA, retrieval often requires distinguishing a relevant table from dense clusters of competing table chunks, and removing those competitors substantially reduces compression-induced degradation. While this analysis is limited to a single dataset and retriever, it points to local competitive structure as an important factor that is not captured by existing corpus-level geometric descriptors.

Shared low-dimensional structure appears sufficient to support strong cross-corpus transfer, yet retrieval robustness ultimately depends on preserving the local distinctions that separate relevant evidence from nearby competitors. Understanding how these local ranking constraints interact with dimensionality reduction remains an important direction for future work.

Beyond the empirical findings, the observed transferability has several practical implications. In settings where a new financial corpus is too small to support reliable PCA fitting, a projection learned on a larger financial collection may provide an effective compression basis without corpus-specific retraining. Transferable projections may also simplify deployment in environments where direct access to corpus-wide document embeddings

is restricted by confidentiality, compliance, or data governance requirements. More broadly, the results suggest that compression bases can sometimes be treated as reusable components across related financial retrieval collections, reducing the operational cost of vector stores in retrieval systems.

6 Conclusion

We presented an empirical study of PCA-based dimensionality reduction for financial retrieval across three benchmarks and eight dense retrievers. We find that PCA projections transfer surprisingly well across financial corpora, with transfer quality closely related to principal-subspace overlap, consistent with shared low-dimensional structure. We further show that global geometric measures such as intrinsic dimension explain representation redundancy but only partially explain retrieval robustness under dimensionality reduction. Through a case study on DocFinQA, we provide evidence that compression failures appear to be associated with dense clusters of highly similar table chunks, highlighting the importance of local retrieval structure.

Limitations

Our study evaluates PCA transfer only within the financial domain using three retrieval benchmarks. While the results suggest substantial shared low-dimensional structure across financial corpora, we do not establish whether the same transferability extends to retrieval datasets from other domains. Furthermore, the observed relationship between principal-subspace overlap and transfer performance is empirical; although the association is strong and robust across retrievers, the experiments do not establish subspace alignment as the sole determinant of transferability.

Our experiments focus exclusively on PCA-based compression. Whether the same transferability and failure patterns arise for learned compression methods or other dimensionality reduction techniques is not studied.

Our analysis of compression failures is based on a DocFinQA case study using a single retriever (`text-embedding-3-small`). The table-removal intervention provides evidence that competition among highly similar table chunks contributes to retrieval degradation under dimensionality reduction. We do not establish whether the same mechanism explains compression failures across

other datasets, retrieval settings, or retrievers.

Finally, DocFinQA retrieval labels are constructed by aligning FinQA evidence spans to retrieval chunks. Sensitivity analyses indicate that the main conclusions are robust to the alignment procedure, although some label noise is unavoidable.

Disclaimer

The views reflected in this article are the views of the authors and do not necessarily reflect the views of the global EY organization or its member firms.

References

- Hervé Abdi and Lynne J. Williams. 2010. [Principal component analysis](#). *WIREs Computational Statistics*, 2(4):433–459.
- Mira Ait-Saada and Mohamed Nadif. 2023. [Is anisotropy truly harmful? a case study on text clustering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1194–1203, Toronto, Canada. Association for Computational Linguistics.
- Jonathan Bac, Evgeny M. Mirkes, Alexander N. Gorban, Ivan Tyukin, and Andrei Zinovyev. 2021. Scikit-dimension: A Python package for intrinsic dimension estimation. *Entropy*, 23(10):1368.
- 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. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of EMNLP*.
- Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of EMNLP-IJCNLP*.
- Elena Facco, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1):12140.
- Ali Forooghi, Shaghayegh Sadeghi, and Jianguo Lu. 2024. [Whitening not recommended for classification tasks in LLMs](#). In *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*, pages 285–289, Bangkok, Thailand. Association for Computational Linguistics.
- Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. [WhiteningBERT: An easy unsupervised sentence embedding approach](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. [Financebench: A new benchmark for financial question answering](#). *Preprint*, arXiv:2311.11944.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*.
- Arman Khaledian, Amirreza Ghadiridehkordi, and Nari-man Khaledian. 2025. [Pca-rag: Principal component analysis for efficient retrieval-augmented generation](#). *Preprint*, arXiv:2504.08386.
- Sejong Kim, Hyunseo Song, Hyunwoo Seo, and Hyunjun Kim. 2025. [Optimizing retrieval strategies for financial question answering documents in retrieval-augmented generation systems](#). *Preprint*, arXiv:2503.15191.
- Amine Kobeissi and Philippe Langlais. 2026. [Decomposing retrieval failures in rag for long-document financial question answering](#). *Preprint*, arXiv:2602.17981.
- Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Elizaveta Levina and Peter J. Bickel. 2004. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 17.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yongkang Li, Panagiotis Eustratiadis, and Evangelos Kanoulas. 2026. [Spectral tempering for embedding compression in dense passage retrieval](#). In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 3941–3946. ACM.
- Zhenghao Liu, Han Zhang, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Xiaohua Li. 2022. [Dimension reduction for efficient dense retrieval via conditional autoencoder](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5692–5698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Xueguang Ma, Minghan Li, Kai Sun, Ji Xin, and Jimmy Lin. 2021. [Simple and effective unsupervised redundancy elimination to compress dense vectors for](#)

- passage retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2854–2859, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. [Www’18 open challenge: Financial opinion mining and question answering](#). In *Companion Proceedings of the The Web Conference 2018*, WWW ’18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Charles Moslonka, Amaury de Vitry, Arthur Garnier, Hicham Randrianarivo, and Emmanuel Malherbe. 2026. [Ledger: A long-context benchmark of corporate annual reports for grounded financial retrieval and extraction](#). *Preprint*, arXiv:2606.13100.
- Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. [Mteb: Massive text embedding benchmark](#). *Preprint*, arXiv:2210.07316.
- Karmvir Singh Phogat, Chetan Harsha, Sridhar Dasaratha, Shashishekar Ramakrishna, and Sai Akhil Puranam. 2023. [Zero-shot question answering over financial documents using large language models](#). *Preprint*, arXiv:2311.14722.
- Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. [DocFinQA: A long-context financial reasoning dataset](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 445–458, Bangkok, Thailand. Association for Computational Linguistics.
- Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.
- Siqi Sun, Ben Peng Wu, Mali Jin, Peizhen Bai, Hanpei Zhang, and Xingyi Song. 2026. [Esg-bench: Benchmarking long-context esg reports for hallucination mitigation](#). *Preprint*, arXiv:2603.13154.
- Sotaro Takeshita, Yurina Takeshita, Daniel Ruffinelli, and Simone Paolo Ponzetto. 2025. [Randomly removing 50% of dimensions in text embeddings has minimal impact on retrieval and classification tasks](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 27705–27726, Suzhou, China. Association for Computational Linguistics.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. [Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models](#). *Preprint*, arXiv:2104.08663.
- Hayato Tsukagoshi and Ryohei Sasano. 2025. [Redundancy, isotropy, and intrinsic dimensionality of prompt-based text embeddings](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25915–25930, Vienna, Austria. Association for Computational Linguistics.
- Yi-Cheng Wang, Wei-An Wang, and Chu-Song Chen. 2026. [Document-level numerical reasoning across single and multiple tables in financial reports](#). *Preprint*, arXiv:2604.03664.
- Vilém Zouhar, Marius Mosbach, Miaoran Zhang, and Dietrich Klakow. 2022. [Knowledge base index compression via dimensionality and precision reduction](#). In *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*, pages 41–53, Dublin, Ireland and Online. Association for Computational Linguistics.
- Chunsheng Zuo and Daniel Khashabi. 2026. [More than efficiency: Embedding compression improves domain adaptation in dense retrieval](#). *Preprint*, arXiv:2601.13525.

A Dataset and Retrieval Construction Details

This appendix provides additional dataset construction and evaluation details.

FiQA. FiQA (Maia et al., 2018) contains 648 financial queries and 57,638 candidate documents. We follow the standard evaluation protocol and use the full benchmark without modification. PCA is fit directly on corpus document embeddings. Since FiQA does not provide a separate corpus for in-distribution projection fitting, evaluations are performed using the full collection.

DocFinQA. DocFinQA (Reddy et al., 2024) extends FinQA (Chen et al., 2021) from short expert-selected evidence to full SEC filings. The benchmark contains 7,437 question-answer pairs derived from 801 filings, with an average document length of approximately 123k words. Retrieval is formulated as identifying the answer-bearing evidence chunk associated with each FinQA question.

Documents are segmented using a structure-aware chunker that preserves tables and section boundaries. The maximum chunk length is 1,200 tokens, chosen to approximately match the scale of the original FinQA reference contexts. Table content is standardized and serialized row-wise before chunking. This process produces an average of 273 retrieval chunks per document.

Dataset	Docs	Queries	Avg. Units/Doc
FiQA	57,638	648	1
DocFinQA	500	500	273
ESG-Bench	251	251	77

Table 1: Evaluation corpus statistics.

For evaluation, we use 500 filings from the official test split. In-distribution PCA experiments use an additional 500 filings sampled from the training split for projection fitting.

ESG-Bench. ESG-Bench (Sun et al., 2026) evaluates question answering over long ESG reports. We segment reports at page boundaries and use the released supporting pages as retrieval targets. Questions may have multiple relevant pages, with an average of 1.3 supporting pages per query.

Our experiments use 251 reports. Because ESG-Bench does not provide an official train/test split for retrieval, all reports are used for the primary in-sample evaluation. For in-distribution PCA experiments, we additionally construct a document-level 80/20 split, fitting PCA on the training portion and evaluating on the held-out portion. Reports are segmented at page boundaries, yielding an average of 77 retrieval units per report.

B Gold-Label Confidence Threshold

DocFinQA does not provide chunk-level retrieval annotations. Following Reddy et al. (2024), we construct retrieval targets by aligning each question’s original FinQA evidence span to candidate document chunks using IDF-weighted 4-gram overlap and selecting the highest-scoring chunk.

Because lexical overlap is only an approximate proxy for evidence alignment, some examples admit multiple plausible chunk assignments. To quantify this ambiguity, we compute the relative margin

$$\tau = \frac{s_{(1)} - s_{(2)}}{s_{(1)}}, \quad (1)$$

where $s_{(1)}$ and $s_{(2)}$ are the overlap scores of the highest- and second-highest-scoring chunks. The main experiments retain only examples with $\tau \geq 0.2$, reducing the evaluation set from 500 sampled filings to 417 examples.

The threshold is intended solely to improve label quality by removing cases where the evidence-to-chunk mapping is uncertain. To assess the sensitivity of our findings to this choice, we repeat the evaluation with $\tau \in \{0.0, 0.2, 0.5\}$, corresponding

τ	Drop%	N	R@10		nDCG@10		$\Delta\%_R$
			Full	PCA	Full	PCA	
0.0	0	500	55.8	54.3	38.3	35.4	-2.7
0.2	17	417	58.6	57.2	40.6	37.6	-2.6
0.5	39	305	62.1	60.1	43.5	40.2	-3.3

Table 2: Performance averaged over 8 retrievers at $k/D = 0.10$, evaluated on samples with alignment margin $> \tau$. *Drop%* and *N* are the fraction removed and the number of samples retained, and $\Delta\%_R$ is the better PCA variant relative to the full retriever on Recall@10. Higher τ lifts all systems together and leaves the compression gap essentially unchanged.

to removal rates of 0%, 17%, and 39%, respectively.

Table 2 shows that increasing the threshold improves absolute retrieval performance for both full-dimensional and PCA-reduced systems, consistent with the removal of noisier examples. However, the relative effect of dimensionality reduction remains remarkably stable. The compression gap varies by only 0.7 points across the entire threshold sweep, indicating that the paper’s main conclusions are not sensitive to the particular threshold selected.

As a qualitative validation, we manually inspected a sample of retained examples and verified that the aligned chunks contained sufficient information to answer the corresponding FinQA question. This inspection was used only to assess the quality of the alignment procedure and was not used to tune retrieval models, PCA projections, or threshold values.

C PCA Fitting Settings

Most prior studies fit PCA directly on the evaluation corpus (in-sample). In contrast, the main experiments use an in-distribution setting, where PCA is fit on a separate training corpus and applied unchanged at evaluation.

Figure 8 compares these settings on DocFinQA. The compression curves are nearly identical across the full range of retained dimensions, with no statistically significant differences in nDCG@10. Similar results are obtained on ESG-Bench.

The choice between in-sample and in-distribution fitting therefore has little effect on retrieval performance in our setting, motivating the cross-dataset transfer experiments presented in the main paper.

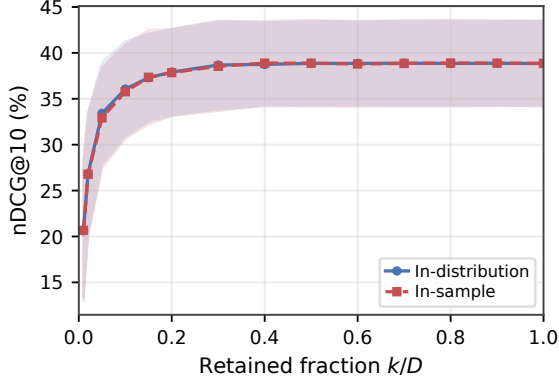


Figure 8: DocFinQA PCA projections fit in-distribution and in-sample produce nearly identical retrieval performance.

Fit unit	N_{fit}	Full	PCA	$\Delta\%$
Page	8,203	79.1	76.8	-2.3
150-token	35,468	79.1	79.9	+0.8
Sentence	100,000	79.1	80.1	+1.0

Table 3: ESG-Bench PCA fit-granularity ablation. Recall@10 at $k/D = 0.10$ for *text-embedding-3-large*. Only the unit used for PCA fitting is varied: entire page, 150 max-token chunk, or single sentence. N_{fit} denotes the number of fitting samples, and $\Delta\%$ reports the relative change compared to the full-dimensional retriever.

D PCA Fit Granularity

For the long-document datasets, PCA is fit on 150 max-token chunks rather than on the retrieval units themselves. This choice raises two questions: (i) does the PCA fitting unit need to match the retrieval unit, and (ii) how sensitive are the results to the fitting granularity?

ESG-Bench. We compare PCA fitted on three granularities derived from the same report collection: report pages (the retrieval unit), 150 max-token chunks, and individual sentences. Table 3 reports Recall@10 at the $k/D = 0.1$.

Both 150 max-token and sentence-level fitting outperform page-level fitting, suggesting that finer-grained text units provide a more useful estimate of the embedding subspace for compression. We attribute this to coarse page representations mixing multiple topics and document-level effects, whereas finer units more directly capture local semantic variation. We therefore fit PCA on 150 max-token chunks throughout the paper; sentence-level fitting yields similar conclusions.

E Robustness of the Transfer–Overlap Correlation

Section 4.2 reports a strong association between source–target principal-subspace overlap and transfer loss (Pearson $r = +0.82$, Spearman $\rho = +0.63$) over $n = 24$ financial transfer settings at $k/D = 0.1$. These observations are not independent: each of the 8 retrievers contributes multiple source–target transfer directions. Consequently, a naive bootstrap that resamples individual points ignores this clustering and yields an overly narrow confidence interval (Pearson $[+0.61, +0.92]$). We therefore apply two clustering-aware robustness checks.

Cluster bootstrap over retrievers. We resample retrievers rather than individual transfer settings, preserving the dependence structure among transfer directions from the same retriever. The resulting 95% confidence interval widens but remains strongly positive (Pearson $[+0.50, +0.93]$). The rank correlation is less stable under this procedure (Spearman $[+0.15, +0.86]$), so we treat the Pearson correlation as the primary estimate.

Leave-one-retriever-out. Recomputing the correlation after removing each retriever in turn yields a consistently strong association: Pearson $[+0.70, +0.86]$ and Spearman $[+0.52, +0.78]$. The weakest Pearson correlation occurs when Granite-s-r2 is excluded, but the relationship remains strong throughout. No single retriever drives the result.

F Centering: Motivation and Analysis

To examine the effect of centering, we vary the centering strength λ from 0 to 1, where $\lambda = 0$ corresponds to uncentered PCA and $\lambda = 1$ to centered PCA. No universal trend is seen in Figure 9. These results suggest treating centering as a model-dependent design choice rather than a fixed preprocessing step.

G Absolute Dimensions per Variance Threshold

Table 4 reports the number of leading principal components required to reach fixed cumulative-variance thresholds on DocFinQA. Although retriever dimensionality varies substantially ($D = 384\text{--}3072$), the absolute number of components required to explain a given fraction of variance

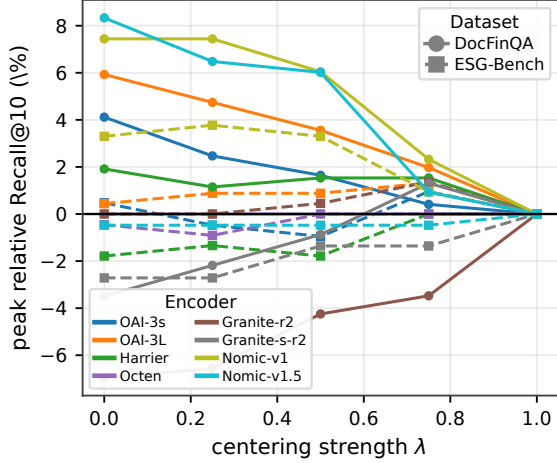


Figure 9: Peak relative Recall@10 as a function of centering strength λ . $\lambda = 0$ corresponds to uncentered PCA and $\lambda = 1$ to centered PCA. No consistent preference emerges across retrievers but difference in performance can be significant in individual cases.

is remarkably stable. For example, reaching 20% variance requires only 6–10 components across all retrievers, while 60% variance requires 61–92 components.

This suggests that the dominant variance structure of the corpus occupies an approximately fixed-dimensional subspace that is largely independent of the ambient embedding dimension. The effect becomes weaker at higher variance thresholds: at 80% variance, the required dimensionality increases from 141 to 264 components, indicating that encoder dimensionality has a greater influence on the lower-variance tail of the spectrum.

We observe the same qualitative behavior on ESG-Bench, despite differences in document type and retrieval task. FiQA exhibits a similar retriever-independent relationship, but requires substantially more components to reach the same variance thresholds. Together, these observations suggest that the absolute number of dimensions needed to capture a fixed fraction of variance is determined primarily by the dataset, while remaining relatively stable across retrievers operating on that dataset.

H Distribution of Query-Local Intrinsic Dimension

Figure 10 shows the distribution of query-local intrinsic dimension using `text-embedding-3-small`. DocFinQA and ESG-Bench queries are anchored in their source document (20 nearest chunks) and FiQA

Encoder	D	Variance threshold			
		20%	40%	60%	80%
OAI-3L	3072	7	27	81	264
OAI-3s	1536	7	25	72	212
Harrier	1024	7	29	92	258
Octen	1024	10	34	87	206
Granite-r2	768	7	27	72	175
Nomic-v1	768	8	31	81	185
Nomic-v1.5	768	6	23	61	148
Granite-s-r2	384	7	25	63	141

Table 4: Number of leading principal components needed by retrievers to reach a fixed cumulative-variance threshold on DocFinQA.

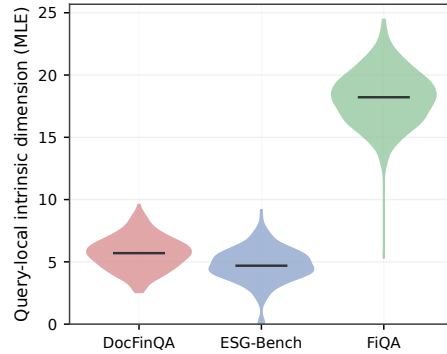


Figure 10: Distribution of query-local intrinsic dimension (Levina–Bickel MLE at full rank, retriever `text-embedding-3-small`) across queries. Horizontal bars mark medians; neighborhood sizes are 20 chunks for the report corpora and 100 for FiQA.

queries in the global corpus (100 nearest chunks). DocFinQA (median 5.7, IQR [4.8, 6.6], $n = 417$) and ESG-Bench (median 4.7, IQR [4.0, 5.6], $n = 251$) overlap almost entirely, even though ESG-Bench compresses markedly better. FiQA occupies a clearly separated, much higher band (median 18.2, IQR [16.7, 19.6], $n = 648$) despite a compression knee comparable to DocFinQA’s.

I Per-Retriever Performance

Figures 11–14 report per-retriever Recall@10 and nDCG@10 as a function of the retained dimension ratio k/D for eight retrievers, each compared with several baselines.

We use the following methods as baselines: **Full** uses the original embeddings without dimensionality reduction. **Matryoshka truncation** retains the first k embedding dimensions followed by L2 nor-

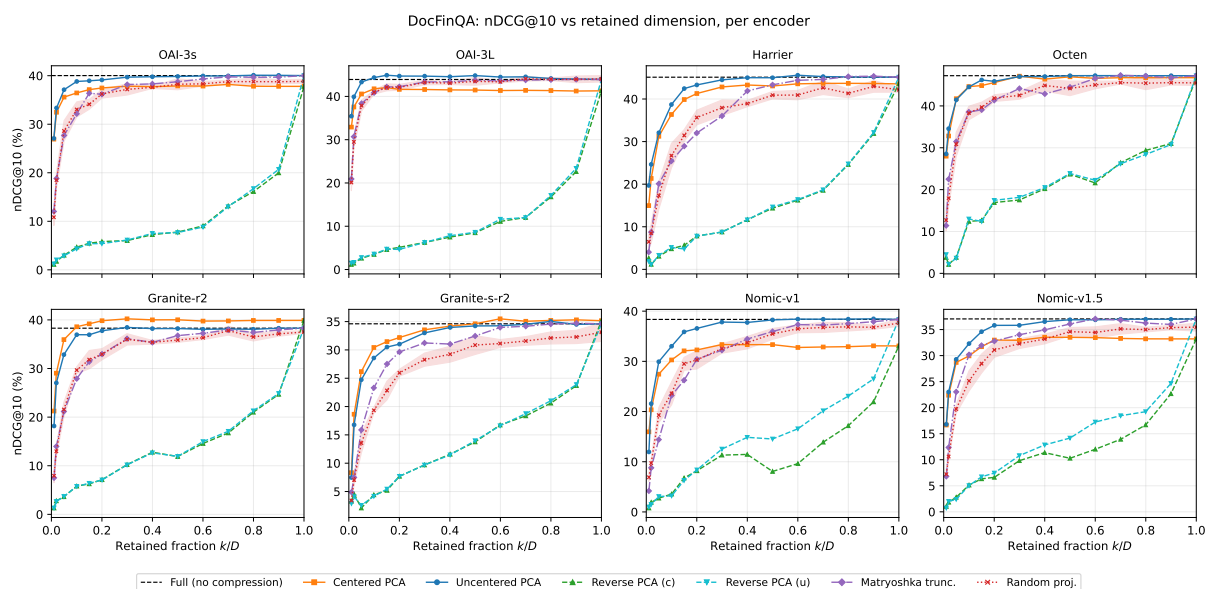
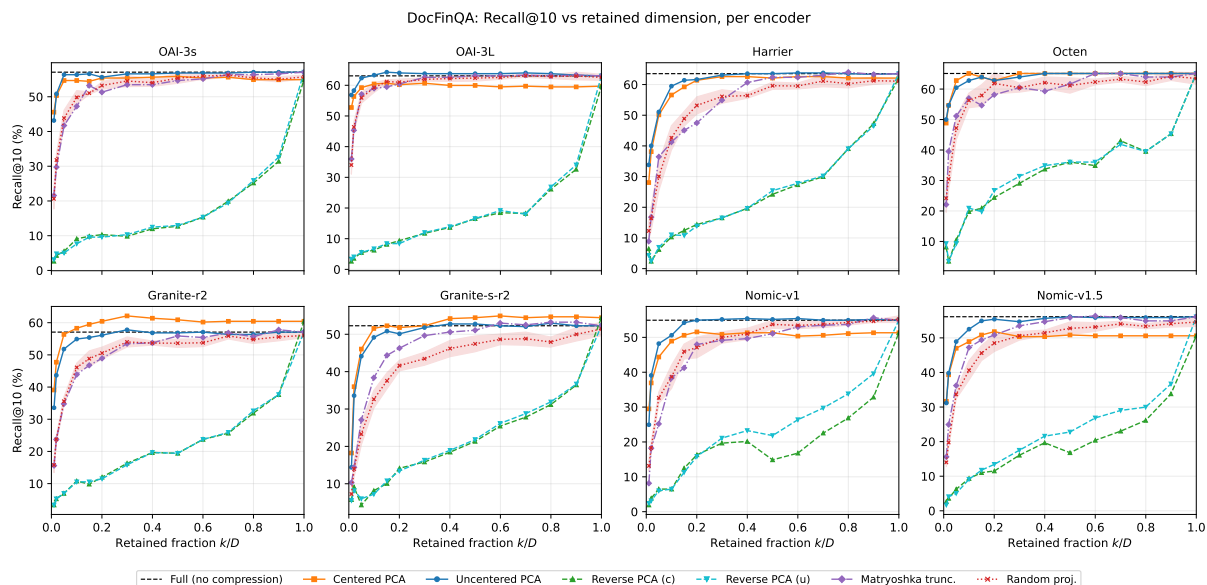
malization. **Random projection** projects embeddings into a randomly selected k -dimensional subspace. **Reverse PCA** projects embeddings onto the bottom k principal components. **Centered PCA** is fit on mean-centered embeddings, whereas **Uncentered PCA** is fit directly on the second moment matrix without mean subtraction.

J Model Details

All models except Nomic-v1 have been trained with Matryoshka representation learning.

Model	Abbrev.	# Params	Dim
text-embedding-3-large	OAI-3L	NA	3072
text-embedding-3-small	OAI-3s	NA	1536
harrier-oss-v1-0.6b	Harrier	600M	1024
Octen-Embedding-0.6B	Octen	600M	1024
nomic-embed-text-v1	Nomic-v1	137M	768
nomic-embed-text-v1.5	Nomic-v1.5	137M	768
granite-english-r2	Granite-r2	149M	768
granite-english-small-r2	Granite-s-r2	47M	384

Table 5: Overview of retrieval models.



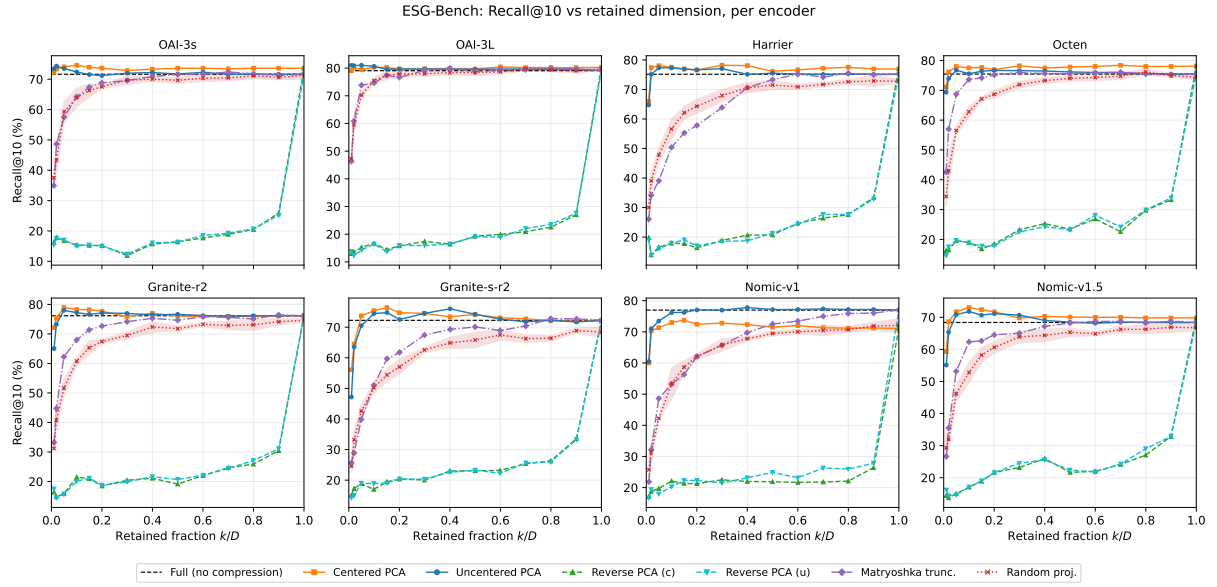


Figure 13: ESG-Bench Recall@10 vs retained dimension k/D (linear axis), per retriever.

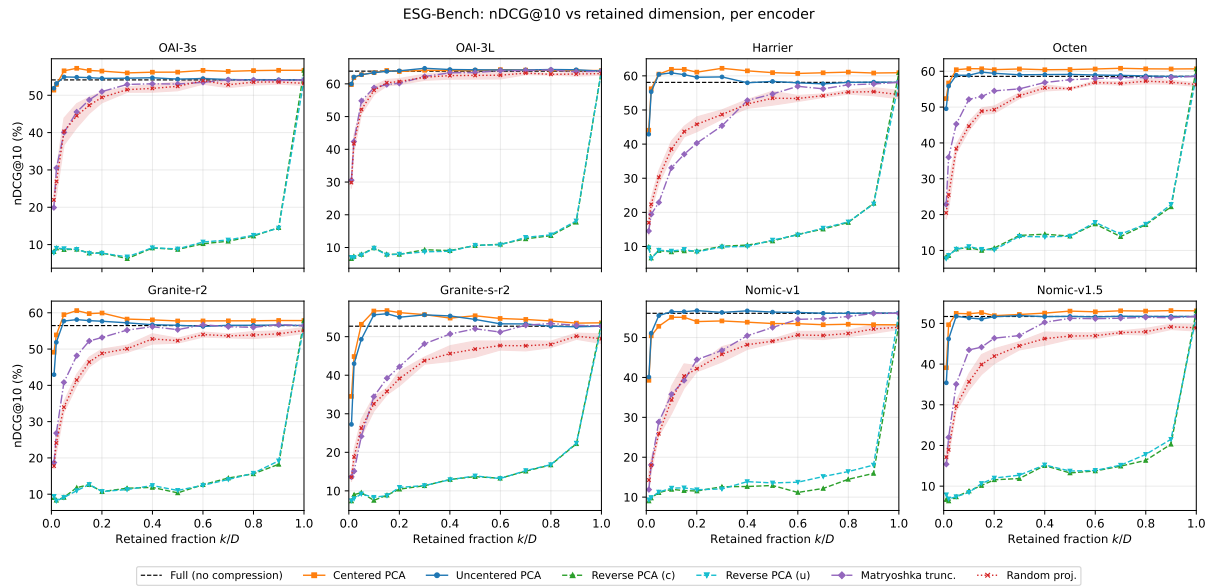


Figure 14: ESG-Bench nDCG@10 vs retained dimension k/D (linear axis), per retriever.