

Lost in Specialization: Embedding Geometry and Retrieval Quality Across Specialized Corpora

Cornelia Ferner^{1,2}, Stefan Wegenkittl¹,

¹Information Technologies and Digitalisation

Salzburg University of Applied Sciences, Austria

²Department of Artificial Intelligence and Human Interfaces

Paris Lodron University of Salzburg, Austria

Correspondence: cornelia.ferner@fh-salzburg.ac.at

Abstract

Reliably operating LLMs in specialized domains often requires retrieval-augmented generation (RAG), which grounds the model’s responses in domain-specific information. Retrieval performance is critical for output quality, making retrieval model selection a key decision for specialized RAG deployments. To guide retrieval model selection, we ask whether domain-tuned encoders produce better-structured embedding spaces for specialized corpora than general-purpose encoders, and whether geometric regularity alone reliably indicates retrieval quality in these settings. Our results show that the answer is more nuanced than expected: domain-specific models show more pronounced geometric irregularity than general-domain encoders, while geometric regularity alone does not reliably indicate retrieval quality. Moreover, geometric and intervention patterns observed on MS-MARCO do not consistently transfer to specialized corpora.

1 Introduction

Retrieval augmented generation (RAG) extends a large language model’s (LLM) global parametric knowledge with domain-specific external information. RAG systems retrieve relevant documents at inference time and inject them into the LLM’s context, enabling factually grounded responses on topics beyond the model’s training data. The quality of this retrieval step is a critical bottleneck, as retrieving wrong documents results in irrelevant or wrong responses from the language model. In practice, many RAG systems are deployed in specialized domains such as medicine, finance or law, where corpora exhibit domain-specific vocabulary, formal register and high topical concentration. The embedding model used in the retrieval step, however, is typically a general-purpose encoder trained on heterogeneous data.

This mismatch between model training distribution and retrieval corpus is well-documented at the

level of retrieval performance: dense retrieval models that achieve strong results on general-purpose benchmarks such as MS-MARCO often degrade substantially on specialized-domain corpora, a finding established systematically by the BeIR benchmark (Thakur et al., 2021). One possible geometric explanation is that domain-specific corpora are embedded in a narrow subspace, a phenomenon referred to as anisotropy. Consequently, if all documents of a specific corpus are projected into roughly the same region of the embedding space, similarity measures lose their discriminative power and nearest-neighbor retrieval degrades.

One natural response is to use encoders pre-trained on domain-specific corpora. A model trained on biomedical text should produce better-structured embeddings for that corpus. Yet domain-specific models are not always trained with retrieval objectives. Many such models are fine-tuned for classification tasks, raising the question of whether domain match alone is sufficient to produce a well-structured embedding space for retrieval. Fine-tuning a general-purpose retrieval model on in-domain data, on the other hand, requires labelled query-document relevance pairs that are scarce or absent in many specialized domains.

In this paper, we study whether geometric properties of the embedding space predict retrieval quality in specialized domains. We quantify geometry using anisotropy, uniformity, effective rank and hubness, and analyze their relationship with retrieval performance (nDCG@10) across biomedical and financial datasets from BEIR, alongside MS-MARCO as a general baseline. As summarized in Table 1, the selected encoders differ along two dimensions: domain specialization and retrieval-oriented training. Our results show more complex patterns than a simple isotropy–performance link and suggest that retrieval-oriented training is an important factor beyond domain match alone.

	Retrieval-oriented training	
	No	Yes
General-domain	SBERT	Contriever BGE
Domain-specific	BioLinkBERT FinBERT	S-PubMedBERT

Table 1: Encoder models considered in this study.

2 Related Work

In RAG systems (Lewis et al., 2020), retrieval quality directly determines downstream LLM generation quality (Shi et al., 2024). While sparse retrieval methods such as BM25 show more robust generalization to out-of-domain corpora (Thakur et al., 2021), dense retrieval encoders are particularly sensitive to domain shift. Our work addresses this domain gap from a geometric perspective, analyzing how embedding-space properties change across specialized corpora and how they relate to retrieval degradation.

2.1 Geometry of Embedding Spaces

Geometric properties of representations learned by language models have first been studied on word level. Ethayarajh (2019) introduced average cosine similarity as a measure of anisotropy, which describes the tendency of representations to occupy only a narrow subspace rather than populating the full embedding space, and finds that representations as learned by BERT and GPT-2 are highly anisotropic. Rudman et al. (2022) introduce IsoScore as a more precise measure of how uniformly the dimensions of the embedding space are utilized and Rudman and Eickhoff (2024) have subsequently shown that decreasing anisotropy can hurt model performance on standard language classification tasks.

Uniformity provides a complementary characterization of embedding space geometry to anisotropy. Wang and Isola (2020) formalize uniformity on the unit hypersphere as a desirable property of learned representations and show that contrastive objectives implicitly optimize for it.

Radovanović et al. (2010) show that in high-dimensional representation spaces, the distribution of k-occurrences becomes skewed, resulting in “popular” representations that occur more frequently among the k-nearest neighbors of other representations. Nielsen and Hansen (2024) document hubness specifically in SBERT embedding spaces, establishing it as a measurable geometric

property of sentence encoders.

Post-hoc correction of geometric irregularity has been explored by Mu and Viswanath (2018) for word embeddings through removal of the dominant PCA directions, and by Jung et al. (2023) who apply isotropy-improving methods in the context of dense passage retrieval, finding score improvements on the target task. We extend this line of work to specialized-domain corpora and treat mean-centering not as a performance optimization but as a diagnostic tool to distinguish embedding spaces where the dominant direction is incidental to retrieval from those where it carries task-relevant signal.

In their comprehensive study, Rottach et al. (2025) conduct a broad topological analysis of text embedding spaces across models and datasets, introducing Unified Topological Signatures (UTS) as a holistic characterization framework and finding that individual geometric measures are highly redundant in general settings. Most importantly, they find effective rank and IsoScore to be the most predictive geometric features for retrieval performance. Our work complements their analysis by examining how these geometric relationships change in specialized-domain settings and whether individual measures remain informative for retrieval model selection.

2.2 Dense Retrieval Models and Benchmark

Dense retrieval encoders evolved from general sentence encoders: With Sentence-BERT (SBERT), Reimers and Gurevych (2019) establish that fine-tuning BERT with a Siamese architecture on sentence pairs produces substantially better embeddings than [CLS] pooling alone. Karpukhin et al. (2020) extend this paradigm to retrieval via a dual-encoder architecture trained on labeled query-passage pairs, demonstrating that dense representations can outperform sparse methods such as BM25. The requirement for labeled pairs limits applicability in specialized domains, motivating unsupervised approaches: With Contriever, Izacard et al. (2022) show that contrastive pretraining without relevance labels achieves competitive zero-shot retrieval. Xiao et al. (2024) scale supervised contrastive retrieval training with large-scale heterogeneous data, producing state-of-the-art general retrieval performance with BGE.

Gao et al. (2021) further connect contrastive training objectives to the geometric properties studied here, showing that in-batch negative training im-

plicitly optimizes for uniformity (Wang and Isola, 2020). As our paper focuses on zero-shot retrieval in specialized domains, we evaluate on the BEIR benchmark (Thakur et al., 2021), which provides a standardized evaluation framework across heterogeneous retrieval corpora.

3 Experimental Setup

We evaluate six encoder models (see Table 1) that differ in domain specialization and retrieval-oriented training. We compare them across four corpora of varying domain specificity, measuring both embedding-space geometry and downstream retrieval performance. No model training or fine-tuning is performed in our experiments; all evaluations use fixed pretrained encoders.

3.1 Models and Datasets

We study the geometry-retrieval link across specialized domains compared with a general-domain corpus. We therefore use MS-MARCO (general, 8.8M passages), TREC-Covid (biomedical, 171K abstracts), NFCorpus (biomedical, 3.6K documents) and FiQA (financial, 57K passages) datasets from the BeIR benchmark¹. The domains represent qualitatively distinct types of specialization: heterogeneous web text (MS-MARCO), dense biomedical terminology in formal scientific register (TREC-Covid, NFCorpus), financial and economic language with the pragmatic register of community forum text (FiQA).

We assess models spanning different combinations of domain specialization and retrieval-oriented training: SBERT², a general-domain sentence encoder optimized for semantic similarity, is our baseline. We compare against two general-domain retrieval-tuned encoders trained with contrastive objectives: Contriever fine-tuned on MS-MARCO³ and BGE⁴, trained on large-scale heterogeneous retrieval data including MS-MARCO. None of these three models has been exposed to domain-specific training data, making them directly comparable as zero-shot retrievers on specialized corpora.

A natural question is whether domain-specific encoders pretrained on in-domain text produce

better-structured embedding spaces for retrieval on specialized corpora, as domain-adaptive pretraining has been shown to improve downstream NLP task performance (Gururangan et al., 2020). To cover different combinations of domain specialization and training objective, we include three different domain-specific encoders. For the biomedical domain, we include BioLinkBERT⁵ (Yasunaga et al., 2022), pre-trained with a citation-link objective and fine-tuned for classification tasks, and S-PubMedBERT⁶, a PubMedBERT-based (Gu et al., 2021) encoder subsequently fine-tuned on MS-MARCO with a semantic similarity objective. For the financial domain, we include FinBERT⁷ (Araci, 2019), fine-tuned for sentiment classification.

Suitable retrieval-trained encoders are not readily available for all specialized domains, while further in-domain fine-tuning requires relevance-labeled query–document pairs that may be scarce or unavailable. We therefore focus on existing pretrained encoders covering several combinations of domain specialization and training objective, rather than training or further adapting models for this study.

All models are bi-encoders, encoding queries and documents independently into single 768-dimensional vectors. Those embeddings are obtained with mean pooling for SBERT, Contriever, BioLinkBERT, S-PubMedBERT, and FinBERT, and CLS pooling for BGE. By choosing the same model dimensionality, we control for the effect of embedding dimensionality on geometric measures, particularly effective rank which was shown to confound cross-model geometry comparisons (Rottach et al., 2025). LLM-based embedding models produce higher-dimensional representations and are excluded for the same reason.

3.2 Geometric Analysis

We measure each model’s embedding-space geometry individually for each corpus.

Anisotropy μ is measured using the mean cosine similarity of embedding vectors (Ethayarajh, 2019) rather than IsoScore (Rudman et al., 2022) for comparability with prior work on retrieval embedding geometry (Jung et al., 2023; Nielsen and Hansen, 2024).

¹<https://huggingface.co/datasets/BeIR>

²<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

³<https://huggingface.co/facebook/contriever-msmarco>

⁴<https://huggingface.co/BAAI/bge-base-en-v1.5>

⁵<https://huggingface.co/michiyasunaga/BioLinkBERT-base>

⁶<https://huggingface.co/pritamdeka/S-PubMedBERT-MS-MARCO>

⁷<https://huggingface.co/ProsusAI/finbert>

Uniformity on the unit hypersphere U is computed on the normalized embedding vectors $\mathbf{z}$ according to Wang and Isola (2020):

$$U = \log\left(\text{mean}_{i \neq j} [\exp(-4 + 4 \cdot \cos(\mathbf{z}_i, \mathbf{z}_j))]\right) \quad (1)$$

Effective rank ER is included as a complementary measure, motivated by Rottach et al. (2025) who find it to be the most predictive geometric feature for retrieval performance in a Random Forest trained on topological signature. Following Roy and Vetterli (2007), effective rank is defined as the Shannon entropy of the normalized singular values of the embedding matrix $\mathbf{X}$:

$$\text{ER}(\mathbf{X}) = \exp\left(-\sum_i p_i \log p_i\right) \quad (2)$$

where $p_i = \sigma_i / \sum_j \sigma_j$ and σ_i are the singular values of $\mathbf{X}$.

Hubness S_k is measured as skewness of the k -occurrence distribution (Radovanović et al., 2010). k -occurrence is defined as the number of times a document is listed in the k nearest neighbors of another document.

S_k is computed on the full corpus, except for MS-MARCO, where the value is approximated by a random sample of size 50 000. We use a FAISS flat inner product index for efficient nearest neighbor search, equivalent to cosine similarity on the unit-normalized embeddings. For μ , U and ER, we compute mean and standard deviation for each measure on three randomly sampled subsets of size 10 000. For NFCorpus, we use the full corpus. Additionally, to assess whether mean embedding directions are corpus-specific or model-induced, we compute the pairwise cosine similarity between the normalized mean vectors of each corpus per model.

3.3 Retrieval Evaluation and Intervention

For the retrieval pipeline, documents, queries and relevance scores are extracted from the BeIR benchmark. For each model-corpus pair, a FAISS flat inner product index is built over the document embeddings and retrieval is performed by querying with the encoded query vectors. Retrieval quality is measured using nDCG@10 (Normalized Discounted Cumulative Gain at rank 10), the standard BeIR evaluation metric which rewards ranking relevant documents higher and penalizes relevant documents appearing lower in the retrieved list. We use

zero-shot evaluation without reranking, query expansion, or retrieval-specific augmentation, as our interest lies in the geometric properties of the base embedding space rather than maximizing retrieval performance.

To probe whether geometric irregularities are incidental or functional for retrieval, we apply the mean-centering step of Mu and Viswanath (2018) as an intervention, without additionally removing top PCA directions. We use this minimal intervention to diagnose whether their findings apply to retrieval, namely that the mean direction is shared across embeddings, or whether this direction instead carries retrieval-relevant information. The corpus mean vector is subtracted from all document and query embeddings and the resulting vectors are renormalized to the unit sphere before retrieval. We report $\Delta\text{nDCG@10} = \text{nDCG@10}_{\text{centered}} - \text{nDCG@10}_{\text{original}}$ per model-corpus pair, where positive values indicate that removing the mean direction improves retrieval.

4 Results

We report geometric properties (see Figure 1a) and retrieval performance (see Figure 2) for the six encoder models across the four corpora, analyzing how embedding space geometry relates to retrieval quality in specialized domains and what the effect of mean-centering as a geometric intervention reveals about the role of the mean embedding direction (see Figure 1b).

Across corpora for which random subsampling is required, the estimates are highly stable, with standard deviations below 0.0016 for anisotropy, 0.0052 for uniformity, and 1.94 for effective rank.

4.1 Geometric Properties across Models and Corpora

Contrary to the hypothesis that domain pretraining produces better-structured embedding spaces for specialized corpora, domain-specific encoders exhibit substantially more severe geometric irregularity than general-domain encoders (see Figure 1a). They exhibit higher anisotropy, lower uniformity, and lower effective rank. No general-domain encoder exceeds $\mu = 0.644$ on any corpus and no domain-specific encoder falls below $\mu = 0.714$. Uniformity mirrors this ordering without exception: Across all corpora, the ranking of models by μ is identical to the ranking by U , suggesting that the two measures are effectively redundant for model

	Anisotropy				Uniformity				Effective Rank				Hubness			
SBERT	0.018	0.186	0.252	0.065	-3.885	-3.025	-2.813	-3.647	554.5	474.8	409.8	513.8	2.170	2.279	1.408	1.941
Contriever	0.238	0.365	0.407	0.354	-3.017	-2.446	-2.320	-2.519	635.5	584.4	531.8	607.7	2.579	5.983	1.425	3.240
BGE	0.386	0.551	0.644	0.549	-2.425	-1.748	-1.385	-1.753	504.6	482.3	434.7	473.0	2.295	2.419	2.061	3.737
S-PubMedBERT	0.857	0.881	0.900	0.885	-0.569	-0.469	-0.397	-0.458	294.6	272.0	225.0	266.5	5.423	3.387	1.920	5.274
BioLinkBERT	0.803	0.809	0.871	0.908	-0.757	-0.717	-0.507	-0.358	354.9	341.7	283.9	266.8	5.085	2.740	2.048	4.736
FinBERT	0.714	0.759	0.858	0.793	-1.103	-0.869	-0.550	-0.786	384.0	332.3	247.8	321.8	3.432	3.006	2.309	3.301
	MS-MARCO	TREC-Covid	NFCorpus	FiQA	MS-MARCO	TREC-Covid	NFCorpus	FiQA	MS-MARCO	TREC-Covid	NFCorpus	FiQA	MS-MARCO	TREC-Covid	NFCorpus	FiQA

(a) Results with original embeddings.

	Anisotropy (centered)				Uniformity (centered)				Effective Rank (centered)				Hubness (centered)			
SBERT	0.000	0.001	0.000	0.000	-3.957	-3.819	-3.817	-3.908	556.0	488.6	431.2	520.0	2.192	2.001	0.970	1.813
Contriever	0.001	0.000	-0.000	0.000	-3.949	-3.889	-3.908	-3.923	662.5	622.2	576.1	645.5	2.326	8.690	1.097	2.337
BGE	0.000	0.000	0.000	0.000	-3.944	-3.872	-3.902	-3.889	544.7	545.8	511.6	535.8	1.689	2.037	0.776	2.138
S-PubMedBERT	0.001	0.001	-0.000	0.000	-3.897	-3.858	-3.875	-3.899	433.5	418.9	377.1	415.9	2.443	3.211	1.010	2.547
BioLinkBERT	0.004	0.003	0.000	0.002	-3.643	-3.557	-3.663	-3.754	490.8	480.8	448.3	461.9	5.840	3.087	2.076	3.708
FinBERT	0.001	0.005	0.001	0.002	-3.794	-3.560	-3.562	-3.713	484.0	435.7	401.7	445.8	2.535	2.602	1.792	2.772
	MS-MARCO	TREC-Covid	NFCorpus	FiQA	MS-MARCO	TREC-Covid	NFCorpus	FiQA	MS-MARCO	TREC-Covid	NFCorpus	FiQA	MS-MARCO	TREC-Covid	NFCorpus	FiQA

(b) Results with mean-centered embeddings.

Figure 1: Geometric properties for all model-corpus combinations: anisotropy (μ), uniformity (U), effective rank (ER) and hubness (S_{10}). For all measures, blue indicates geometrically favorable values (low μ , low U , high ER, low S_{10}) and red indicates unfavorable values.

comparison even in specialized-domain settings.

Effective rank shows the same categorical separation, with domain-specific encoders consistently below general-domain encoders. The highest effective rank among domain-specific encoders is $ER = 384.0$ for FinBERT on MS-MARCO, while the lowest value among general-domain encoders is $ER = 409.8$ for SBERT on NFCorpus. Contriever has the highest effective rank across all corpora.

Hubness values show a different pattern. First, Contriever stands out on TREC-Covid with $S_{10} = 5.983$ which is more than four times its value on NFCorpus (1.425) and far above any other general-domain encoder on any corpus. Second, BGE has higher hubness than corresponding domain-specific models on matched specialized corpora: On FiQA, BGE’s hubness ($S_{10} = 3.737$) exceeds FinBERT’s (3.301) and on NFCorpus it exceeds S-PubMedBERT (BGE 2.061 vs. 1.920).

Within the domain-specific group, BioLinkBERT shows a somewhat more regular embedding space than the retrieval-tuned S-PubMedBERT on their biomedical target corpus TREC-Covid: lower anisotropy ($\mu = 0.809$ vs

0.881), higher effective rank ($ER = 341.7$ vs 272.0), better uniformity ($U = -0.717$ vs -0.469) and lower hubness ($S_{10} = 2.740$ vs 3.387). By the geometric reasoning that motivates our analysis, these measures would favor BioLinkBERT as the stronger retrieval model.

Subsequently, in Table 2 we analyze whether the geometric differences across corpora reflect domain-specific directional concentration or a model-induced property by computing the pairwise cosine similarity between the normalized mean vectors per corpus. SBERT’s mean directions are nearly orthogonal across the two specialized corpora ($FiQA \leftrightarrow TREC-Covid$ 0.11), indicating that each domain pulls embeddings into a distinct subspace which can be considered corpus-driven anisotropy.

By contrast, S-PubMedBERT’s mean direction is essentially identical regardless of corpus (0.97–0.99), suggesting a largely model-induced property that is relatively insensitive to corpus content. Notably, Contriever shows the largest gap between its MS-MARCO $\leftrightarrow$ TREC-Covid (0.77) and MS-MARCO $\leftrightarrow$ FiQA (0.86) alignment val-

	T $\leftrightarrow$ F	M $\leftrightarrow$ T	M $\leftrightarrow$ F
SBERT	0.11	0.29	0.32
Contriever	0.65	0.77	0.86
BGE	0.78	0.88	0.90
FinBERT	0.88	0.93	0.97
BioLinkBERT	0.86	0.94	0.97
S-PubMedBERT	0.97	0.98	0.99

Table 2: Pairwise cosine similarity of corpus mean vectors per model: TREC-Covid vs. FiQA (T $\leftrightarrow$ F), MS-MARCO vs. TREC-Covid (M $\leftrightarrow$ T) and MS-MARCO vs. FiQA (M $\leftrightarrow$ F).

ues of all models. This corpus-specific shift coincides with its exceptionally high hubness on TREC-Covid reported above.

4.2 Geometry vs. Retrieval Quality

Retrieval performance in Figure 2 (left) shows a different pattern from the geometric separation. General-domain encoders achieve nDCG@10 between 0.274 and 0.713, whereas the domain-specific models without retrieval-oriented training (BioLinkBERT and FinBERT) collapse to 0.010 and 0.025, respectively. BGE is the strongest retrieval model overall with nDCG@10=0.713 on MS-MARCO and 0.672 on TREC-Covid.

S-PubMedBERT, the only domain-specific and retrieval-tuned model, achieves competitive retrieval performance on TREC-Covid with nDCG@10=0.592 compared to SBERT with 0.513 and BGE with 0.672. This is surprising given the high anisotropy (0.881) and lowest ER of any model (272). On the same corpus, BioLinkBERT achieves nDCG@10=0.129 with $\mu = 0.809$ and ER = 341.7. Thus, S-PubMedBERT is more anisotropic and uses fewer dimensions yet retrieves 4.6 times better. **None of the four geometric measures distinguishes BioLinkBERT’s poor retrieval performance from S-PubMedBERT’s competitive performance.**

Most models, except for BioLinkBERT and FinBERT, have their strongest retrieval performance and healthiest geometric measures on MS-MARCO. **However, favorable geometric profiles observed on MS-MARCO do not consistently transfer to specialized corpora, nor do they reliably indicate retrieval performance.** Contriever, for example, exhibits exceptionally high hubness on TREC-Covid, while SBERT shows comparatively high anisotropy, low uniformity, and low ef-

fective rank on NFCorpus, alongside substantially lower nDCG@10 than on its other evaluated corpora.

A well-structured embedding space is not, by itself, a reliable indicator of retrieval performance. BGE achieves the strongest retrieval performance without exhibiting the most favorable geometric values across all measures. Conversely, BioLinkBERT’s higher effective dimensionality compared with S-PubMedBERT does not correspond to better retrieval performance. Extending the observations of Rottach et al. (2025), our results show that individual geometric measures alone do not provide a reliable criterion for selecting retrieval encoders in specialized-domain settings.

4.3 Mean-centering Intervention

We apply mean-centering as a diagnostic tool to probe the role of the mean direction in each model’s embedding space. The intervention allows us to assess whether this direction carries signal relevant for retrieval ranking or is incidental.

Figure 1b illustrates the geometric metrics after the intervention. Mean-centering reduces anisotropy to near-zero and uniformity to below -3.5 for all models and corpora, showing that both measures are strongly driven by the mean direction. Effective rank increases after centering for all models: substantially so for domain-specific encoders (BioLinkBERT TREC-Covid: ER=341.7 $\rightarrow$ 480.8, S-PubMedBERT TREC-Covid: 272.0 $\rightarrow$ 418.9) and modestly for general-domain encoders (SBERT TREC-Covid: 474.8 $\rightarrow$ 488.6). Hubness generally decreases after centering, although several model-corpus combinations show increases, most notably Contriever on TREC-Covid (S_{10} =5.983 $\rightarrow$ 8.690) and BioLinkBERT on MS-MARCO (S_{10} =5.085 $\rightarrow$ 5.840).

As Figure 2 (right) illustrates, mean-centering consistently worsens retrieval across all corpora for domain-specific models without retrieval-training. BioLinkBERT shows the largest decrease on TREC-Covid (Δ =-0.110), followed by FinBERT on the same corpus (Δ =-0.065). This consistent degradation indicates that the mean direction in these models carries information relevant to retrieval ranking rather than being merely incidental.

The effect of mean-centering on MS-MARCO does not consistently transfer to specialized corpora. For SBERT and BGE, centering marginally hurts MS-MARCO retrieval (Δ =-0.006 and Δ =-0.026, respectively) but improves retrieval on

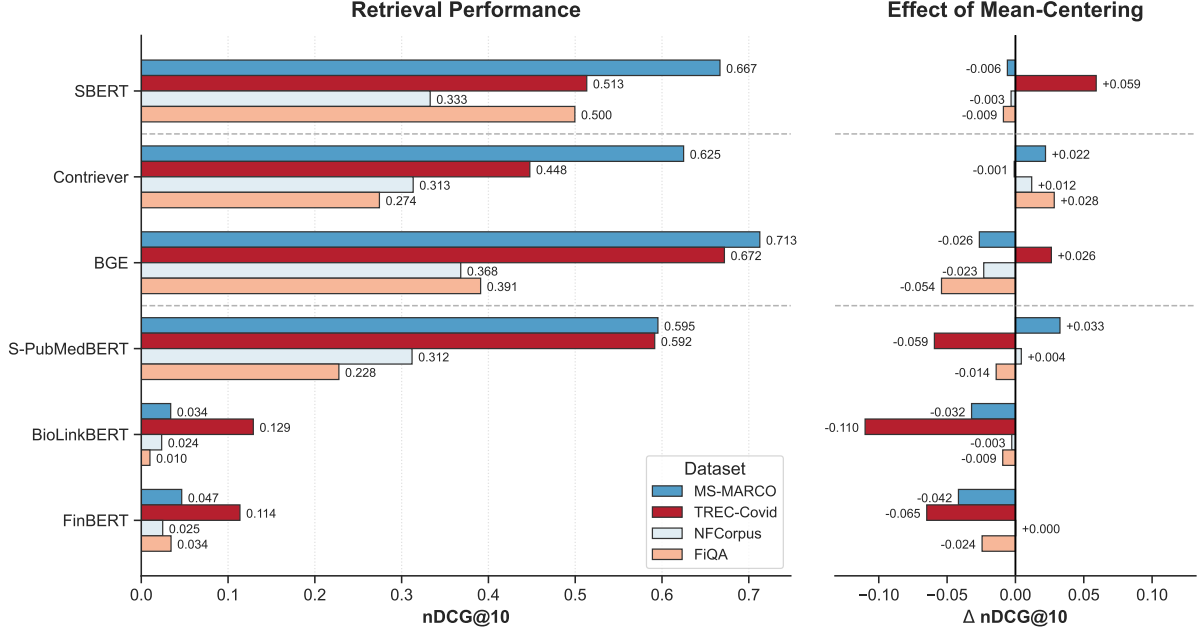


Figure 2: Retrieval performance as measured by nDCG@10 (left) and difference obtained after mean-centering shown as Δ nDCG@10 (right) for all model-corpus combinations.

TREC-Covid ($\Delta=+0.059$ and $\Delta=+0.026$). The contribution of the mean direction to retrieval differs between the general and specialized corpus. For Contriever and S-PubMedBERT, the pattern reverses: centering improves MS-MARCO retrieval ($\Delta=+0.022$ and $\Delta=+0.033$) but has negligible or negative effect on TREC-Covid ($\Delta=-0.001$ and $\Delta=-0.059$). MS-MARCO intervention behavior would therefore be misleading as a model-selection signal for the specialized corpus.

The cross-domain alignment values in Table 2 help contextualize this pattern. SBERT’s low alignment (FiQA $\leftrightarrow$ TREC-Covid: 0.11) indicates strongly corpus-specific mean directions. This is consistent with the corpus-dependent effect of centering: it improves retrieval on TREC-Covid while having little effect on MS-MARCO, where anisotropy is already low. BGE and Contriever show substantially higher cross-domain alignment (0.78 and 0.65), indicating more stable mean directions across corpora. Nevertheless, the retrieval effect of removing these directions still varies by model and corpus, showing that cross-domain directional stability alone does not determine whether the mean direction is beneficial for retrieval.

For S-PubMedBERT, mean-centering improves retrieval on MS-MARCO ($\Delta=+0.033$) but degrades it on TREC-Covid ($\Delta=-0.059$). Its mean direction is exceptionally stable across corpora (pairwise cosine similarity ≥ 0.97), suggesting a largely model-

induced structure. Nevertheless, the opposite intervention effects indicate that the retrieval relevance of this direction depends on the corpus. This is consistent with biomedical pretraining and subsequent MS-MARCO retrieval tuning shaping different aspects of the embedding space.

5 Conclusion

In this paper, we investigated the relationship between geometric properties of retrieval models’ embedding spaces and their performance in specialized domains (biomedicine and finance). Our results show that favorable geometric properties observed on a general-domain corpus (MS-MARCO) do not consistently transfer to specialized corpora and do not reliably indicate retrieval performance there. More generally, geometric regularity alone does not provide a reliable criterion for selecting retrieval encoders on a given specialized corpus. On TREC-Covid, for example, S-PubMedBERT achieves overall competitive retrieval despite a more irregular embedding space than BioLinkBERT: none of the four assessed geometric measures identifies S-PubMedBERT as the substantially stronger retrieval model. Mean-centering further supports this finding. Although it consistently improves several geometric measures, its effect on retrieval varies in both magnitude and direction across models and corpora. In particu-

lar, intervention effects observed on MS-MARCO do not consistently indicate their effects on TREC-Covid.

Our results also highlight that domain match alone is insufficient for model selection and that retrieval-oriented training is an important factor beyond domain-specific pretraining. BioLinkBERT and FinBERT, both domain-specific models without retrieval-oriented training, substantially underperform the retrieval-tuned encoders, whereas S-PubMedBERT achieves competitive retrieval on its biomedical target corpus despite pronounced geometric irregularity. Practitioners should therefore not select retrieval encoders based on domain match or individual geometric properties alone, but also consider whether the model has been trained for retrieval. Rather than assuming that domain-specific pretraining will necessarily yield better retrieval, general-domain retrieval-trained encoders already provide a strong baseline for zero-shot retrieval on a specialized corpus.

Limitations

Our study considers six encoders across four corpora, with specialized-domain evaluation limited to biomedical and financial text. Domain specialization and retrieval-oriented training are not experimentally disentangled, so observed differences should not be attributed to either factor in isolation.

References

- Dogu Araci. 2019. [Finbert: Financial sentiment analysis with pre-trained language models](#). *CoRR*, abs/1908.10063.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. [Domain-specific language model pretraining for biomedical natural language processing](#). *ACM Trans. Comput. Healthcare*, 3(1).
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. [Unsupervised dense information retrieval with contrastive learning](#). *Transactions on Machine Learning Research*.
- Euna Jung, Jungwon Park, Jaekeol Choi, Sungyoon Kim, and Wonjong Rhee. 2023. [Isotropic representation can improve dense retrieval](#). In *Advances in Knowledge Discovery and Data Mining: 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2023, Osaka, Japan, May 25–28, 2023, Proceedings, Part III*, page 125–137, Berlin, Heidelberg. Springer-Verlag.
- Vladimir Karpukhin, Barlas Oguz, 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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA. Curran Associates Inc.
- Jiaqi Mu and Pramod Viswanath. 2018. [All-but-the-top: Simple and effective postprocessing for word representations](#). In *International Conference on Learning Representations*.
- Beatrix M.G. Nielsen and Lars Kai Hansen. 2024. Hubness reduction improves sentence-bert semantic spaces. In *Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL)*, volume 233, pages 181–204. Proceedings of Machine Learning Research. 5th Northern Lights Deep Learning Conference, NLDL ; Conference date: 09-01-2024 Through 11-01-2024.
- Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. [Hubs in space: Popular nearest neighbors in high-dimensional data](#). *Journal of Machine Learning Research*, 11(86):2487–2531.

- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Florian Rottach, William Rudman, Bastian Rieck, Harisen Scells, and Carsten Eickhoff. 2025. [From topology to retrieval: Decoding embedding spaces with unified signatures](#). *Preprint*, arXiv:2511.22150. Preprint.
- Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pages 606–610.
- William Rudman and Carsten Eickhoff. 2024. [Stable anisotropic regularization](#). In *The Twelfth International Conference on Learning Representations*.
- William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. [IsoScore: Measuring the uniformity of embedding space utilization](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339, Dublin, Ireland. Association for Computational Linguistics.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. [REPLUG: Retrieval-augmented black-box language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. [BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models](#). In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML’20*. JMLR.org.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-pack: Packed resources for general chinese embeddings](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’24*, page 641–649, New York, NY, USA. Association for Computing Machinery.
- Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. [LinkBERT: Pretraining language models with document links](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.