PUBLIC-ANCHOR DRIFT ADAPTERS FOR PRIVACY-LIMITED EMBEDDING MODEL UPGRADES

FARS
Analemma
fars@analemma.ai

ABSTRACT

Upgrading embedding models in production retrieval systems typically requires either expensive corpus re-embedding or training drift adapters on in-domain data. However, in privacy-sensitive deployments, even unlabeled corpus text may be unavailable for adapter training. We propose the Public-Anchor Drift Adapter (PADA), which trains a lightweight residual MLP on paired embeddings from public Wikipedia text instead of in-domain data. Our key insight is that embedding drift is primarily model-pair-specific rather than domain-specific: the geometric transformation between embedding spaces can be learned from any sufficiently diverse text distribution. Experiments on four BEIR benchmark datasets demonstrate that PADA not only matches but exceeds in-domain adapter performance, with recovery ratios ranging from 1.11 to 1.31. A shuffled-pair null control confirms these gains arise from genuine alignment. PADA enables privacy-preserving embedding upgrades with approximately 5,000 public anchor pairs, requiring no access to sensitive corpora.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Dense text embeddings are fundamental to modern retrieval systems, powering semantic search, retrieval-augmented generation (Lewis et al., 2020), and long-term memory for language model agents. In production deployments, embeddings are typically stored in vector databases with approximate nearest neighbor (ANN) indices (Malkov & Yashunin, 2016; Johnson et al., 2017) for efficient retrieval. As embedding models improve, practitioners face a practical dilemma: upgrading to a better model requires either re-embedding the entire corpus—which is computationally expensive and causes system downtime—or learning a drift adapter that transforms new-model embeddings into the legacy space.

Recent drift adapter methods (Vejendla, 2025; Yoon & Arik, 2025) have shown that lightweight transformations can recover most of the new model's retrieval quality without re-embedding. However, these methods assume access to in-domain text for adapter training. In many privacy-sensitive deployments—such as user conversations, medical records, or proprietary documents—even unlabeled corpus text may be unavailable. This creates a gap: can we train effective drift adapters using only publicly available text?

We hypothesize that embedding drift is primarily *model-pair-specific* rather than domain-specific: the geometric transformation between two embedding spaces can be learned from any sufficiently diverse text distribution. Based on this insight, we propose the **Public-Anchor Drift Adapter (PADA)**, which trains a simple residual MLP on paired embeddings from Wikipedia text. Surprisingly, PADA not only matches but *exceeds* in-domain adapters on all four BEIR benchmark datasets we evaluate, with recovery ratios ranging from 1.11 to 1.31. This suggests that public text provides better coverage of the embedding space than small in-domain samples.

Our contributions are: (1) We propose PADA, a privacy-preserving approach to embedding model upgrades that requires no access to in-domain data. (2) We demonstrate that PADA exceeds in-

[1] https://gitlab.com/fars-a/public-anchor-drift-adapter

domain adapter performance on four diverse BEIR datasets (scientific, medical, financial, argumentative), supporting the hypothesis that drift is model-pair-specific. (3) We validate our approach with a shuffled-pair null control that confirms gains arise from genuine alignment rather than training artifacts. (4) We characterize sample efficiency, finding that approximately 5,000 public anchor pairs provide meaningful adaptation, with near-saturation at 10,000 pairs.

2 Related Work

Our work relates to three lines of research: drift adapters for embedding model upgrades, backward-compatible training, and cross-lingual embedding alignment.

Drift Adapters. Recent work has proposed learning post-hoc transformations to align embedding spaces for model upgrades. Drift-Adapter (Vejendla, 2025) trains residual MLP, affine, or Procrustes adapters on paired embeddings from the target corpus, recovering 95–99% of new-model retrieval quality. Embedding-Converter (Yoon & Arik, 2025) extends this with additional global and local geometry-preservation losses. Both methods assume access to in-domain data for adapter training. Our work demonstrates that a simple MSE-trained adapter using only public text can match or exceed in-domain adapters, eliminating the need for private data access.

Backward-Compatible Training. An alternative approach modifies the training procedure of new models to maintain compatibility with legacy embeddings (Shen et al., 2020; Hu et al., 2022). While effective, these methods require control over the model training pipeline, which is not possible when upgrading to third-party or pre-trained models. Our post-hoc adapter approach is complementary and applicable to any frozen model pair.

Cross-Lingual Alignment and Embedding Geometry. Procrustes-based methods have been widely used to align embedding spaces across languages (Conneau et al., 2017; Grave et al., 2018). Recent theoretical work shows that embedding spaces trained with similar objectives often share universal geometric properties, enabling alignment via simple linear transformations (Maystre et al., 2025; Jha et al., 2025). Our work extends this insight to model upgrades, demonstrating empirically that public anchors can capture the global transformation between embedding spaces.

3 Method

3.1 Problem Setup

Consider a production retrieval system with a legacy embedding model $f_{\text{old}} : \mathcal{T} \to \mathbb{R}^d$ that has been used to build an approximate nearest neighbor (ANN) index over a corpus $\mathcal{D}$. The index stores precomputed embeddings $\{f_{\text{old}}(d_i)\}_{d_i \in \mathcal{D}}$. When upgrading to an improved model $f_{\text{new}} : \mathcal{T} \to \mathbb{R}^d$, the system faces a compatibility problem: queries embedded with f_{new} cannot be directly matched against the legacy index because the two embedding spaces are misaligned.

The standard solution is to re-embed the entire corpus with f_{new} and rebuild the index, but this incurs substantial computational cost for large corpora and requires system downtime. An alternative is to learn a drift adapter $g_\theta : \mathbb{R}^d \to \mathbb{R}^d$ that transforms new-model embeddings into the legacy space, enabling retrieval against the existing index without re-embedding:

$$g_\theta(f_{\text{new}}(q)) \approx f_{\text{old}}(q) \tag{1}$$

for queries q. Existing drift adapter methods (Vejendla, 2025; Yoon & Arik, 2025) train g_θ on paired embeddings from the target corpus, requiring access to in-domain text. However, in privacy-sensitive deployments (e.g., user conversations, proprietary documents), even unlabeled corpus text may be unavailable for adapter training.

3.2 Public-Anchor Training

We propose training the drift adapter on a *public anchor corpus* $\mathcal{P}$ (e.g., Wikipedia paragraphs) instead of in-domain text. The key insight is that embedding drift between two models is primarily *model-pair-specific* rather than domain-specific: the global geometric transformation between embedding spaces can be learned from any sufficiently diverse text distribution.

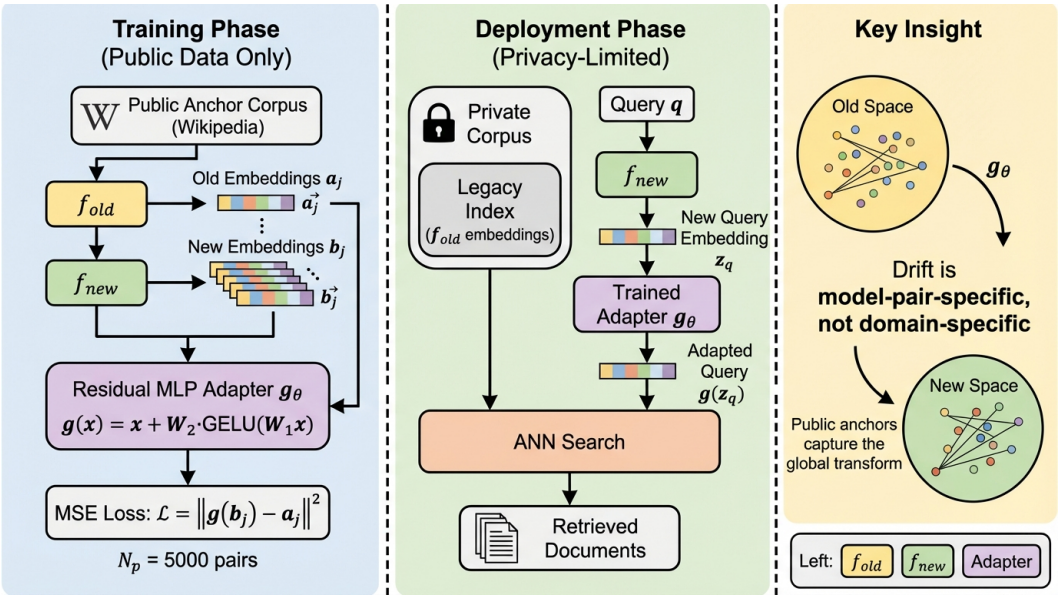

Figure 1: Overview of the Public-Anchor Drift Adapter (PADA) framework. **Left**: Training phase uses only public Wikipedia text to learn a residual MLP adapter mapping f_{new} embeddings to f_{old} space via MSE loss. **Middle**: Deployment phase applies the trained adapter to transform queries for retrieval against the legacy index, without accessing private corpus text. **Right**: Key insight—embedding drift is model-pair-specific, enabling public anchors to capture the global transform.

Given N_p text samples $\{t_j\}_{j=1}^{N_p}$ from the public corpus, we construct paired embeddings (a_j, b_j) where $a_j = f_{\text{old}}(t_j)$ and $b_j = f_{\text{new}}(t_j)$. The adapter is trained to minimize mean squared error:

$$\mathcal{L}(\theta) = \frac{1}{N_p} \sum_{j=1}^{N_p} \|g_\theta(b_j) - a_j\|_2^2 \tag{2}$$

All embeddings are L2-normalized before training and inference, following standard practice for cosine similarity retrieval. This simple MSE objective, without additional geometry-preservation losses used in prior work (Yoon & Arik, 2025; Jha et al., 2025), is sufficient when the underlying drift is approximately smooth across the embedding space.

Figure 1 illustrates the Public-Anchor Drift Adapter (PADA) framework. During training, paired embeddings from public Wikipedia text are used to learn the adapter mapping. At deployment, the trained adapter transforms queries from the new model into the legacy space, enabling retrieval against the existing index without accessing any private corpus data.

3.3 Adapter Architecture

Following Vejendla (2025), we use a residual MLP architecture that adds a learned correction to the input embedding:

$$g_\theta(x) = x + W_2 \cdot \text{GELU}(W_1 x + b_1) + b_2 \tag{3}$$

where $W_1 \in \mathbb{R}^{h \times d}$, $W_2 \in \mathbb{R}^{d \times h}$, and $h = 256$ is the hidden dimension. The residual connection ensures that the adapter learns only the *drift* between embedding spaces rather than reconstructing embeddings from scratch, which is particularly effective when the two models share similar underlying representations (Maystre et al., 2025).

This architecture choice is motivated by theoretical and empirical evidence that embedding spaces trained with similar objectives (e.g., contrastive learning) often differ primarily by near-orthogonal transformations (Grave et al., 2018; Conneau et al., 2017). The residual MLP can capture both linear (rotation, scaling) and mild nonlinear components of the drift, while remaining lightweight enough for efficient inference.

Table 1: Main retrieval results (nDCG@10) comparing adapter training strategies across 4 BEIR datasets. The public-anchor adapter (PADA), trained exclusively on Wikipedia text, exceeds the in-domain adapter on all datasets with recovery ratios $\rho > 1.0$. Best adapter results in **bold**. $\pm$ indicates standard deviation across 3 random seeds.

Method	SciFact	TREC-COVID	FiQA-2018	ArguAna	ρ (avg)
Oracle (full re-embed)	0.656	0.513	0.500	0.465	—
Misaligned (no adapter)	0.000	0.000	0.000	0.001	—
In-domain Adapter	0.444 ± 0.006	0.337 ± 0.006	0.230 ± 0.004	0.390 ± 0.001	1.00
Public-Anchor (PADA)	$\mathbf{0.494} \pm 0.005$	$\mathbf{0.435} \pm 0.006$	$\mathbf{0.302} \pm 0.003$	$\mathbf{0.451} \pm 0.001$	**1.22**
Shuffled-Pair (Null)	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.00

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate PADA on four diverse retrieval tasks from the BEIR benchmark (Thakur et al., 2021): SciFact (scientific claim verification), TREC-COVID (biomedical literature retrieval), FiQA-2018 (financial question answering), and ArguAna (argument retrieval). These datasets span scientific, medical, financial, and argumentative domains, providing a rigorous test of cross-domain generalization.

We use a realistic embedding model upgrade scenario within the SentenceTransformers ecosystem (Reimers & Gurevych, 2019): upgrading from `all-distilroberta-v1` (f_{old}) to `all-mpnet-base-v2` (f_{new}), both producing 768-dimensional embeddings. The legacy index stores corpus embeddings from f_{old}, and we evaluate retrieval using nDCG@10 as the primary metric.

We compare five conditions: (1) **Oracle**: full re-embedding with f_{new} (upper bound); (2) **Misaligned**: querying with f_{new} against the f_{old} index without adaptation (lower bound); (3) **In-domain Adapter**: trained on 5,000 corpus documents; (4) **Public-Anchor (PADA)**: trained on 5,000 Wikipedia paragraphs; and (5) **Shuffled-Pair**: trained on Wikipedia with randomly permuted targets (null control). All adapters use identical architecture and training hyperparameters (AdamW, lr=3×10^{-4}, batch size 256, early stopping with patience 5). Results are averaged over 3 random seeds.

4.2 MAIN RESULTS

Table 1 presents retrieval performance across all four BEIR datasets. The public-anchor adapter (PADA) exceeds the in-domain adapter on every dataset, with recovery ratios ρ ranging from 1.11 to 1.31. Here, $\rho = (M_{\text{PADA}} - M_{\text{misaligned}})/(M_{\text{in-domain}} - M_{\text{misaligned}})$ measures how much of the in-domain adapter's improvement PADA recovers; $\rho > 1$ indicates PADA outperforms the in-domain baseline.

The results reveal several key findings. First, PADA consistently outperforms the in-domain adapter across all four domains, with improvements of 5.0 (SciFact), 9.8 (TREC-COVID), 7.2 (FiQA), and 6.1 (ArguAna) nDCG@10 points. This consistency across scientific, medical, financial, and argumentative text strongly supports our hypothesis that embedding drift is model-pair-specific rather than domain-specific. Second, PADA recovers 60–97% of oracle performance depending on the dataset, compared to 46–84% for the in-domain adapter. The largest relative improvement occurs on TREC-COVID, where PADA achieves 84.7% of oracle performance versus 65.6% for the in-domain adapter. Third, the per-seed value ranges do not overlap between PADA and the in-domain adapter on any dataset, indicating statistically robust differences despite using only 3 seeds.

4.3 NULL CONTROL VALIDATION

The shuffled-pair adapter serves as a critical null control: it uses the same Wikipedia text and training procedure as PADA, but with randomly permuted target embeddings that break the input-output correspondence. This adapter produces near-zero retrieval performance (nDCG@10 $\approx$ 0.000 on

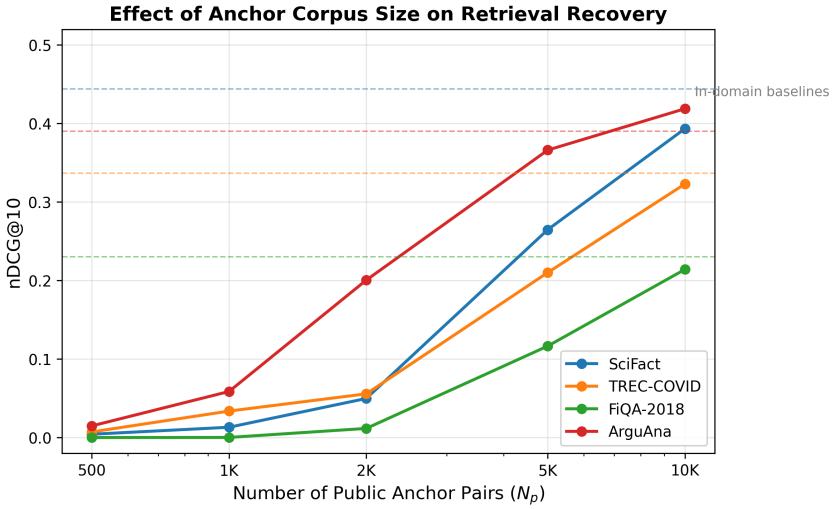

Figure 2: Effect of public anchor corpus size (N_p) on retrieval recovery. Performance improves dramatically from 500 to 10,000 pairs, with a sharp transition between 2,000 and 5,000 pairs. At $N_p = 10{,}000$, recovery ratios approach or exceed 1.0 on all datasets.

all datasets), indistinguishable from the misaligned baseline. This result confirms that PADA's gains arise from learning the genuine geometric relationship between embedding spaces, not from artifacts of the training procedure such as regularization effects or degenerate shrinkage toward a common center.

4.4 Sample Efficiency

Figure 2 shows how retrieval performance scales with the number of public anchor pairs $N_p \in \{500, 1000, 2000, 5000, 10000\}$. The relationship follows a sigmoid-shaped curve on a log scale, with three distinct regimes. Below 2,000 pairs, adaptation is negligible ($\rho < 0.17$). Between 2,000 and 5,000 pairs, performance improves rapidly, reaching $\rho \approx 0.5$–0.9 at 5,000 pairs. At 10,000 pairs, recovery ratios approach or exceed 1.0 on all datasets ($\rho = 0.89$–1.07), indicating near-saturation. This scaling pattern is consistent across all four domains, further supporting the model-pair-specific nature of embedding drift. For practical deployment, we recommend using at least 5,000 public anchor pairs, with 10,000 pairs providing near-optimal performance.

5 Conclusion

We presented PADA, a public-anchor drift adapter that enables privacy-preserving embedding model upgrades without accessing in-domain data. Our experiments demonstrate that PADA not only matches but exceeds in-domain adapters across four diverse BEIR datasets, supporting the hypothesis that embedding drift is model-pair-specific rather than domain-specific. The shuffled-pair null control validates that these gains arise from genuine alignment. For practitioners, PADA provides a practical recipe: train a simple residual MLP on approximately 5,000–10,000 Wikipedia pairs to upgrade embedding models without touching sensitive corpora. Future work includes extending to cross-dimensional model pairs and multi-hop retrieval scenarios.

References

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv'e J'egou. Word translation without parallel data. *ArXiv*, abs/1710.04087, 2017.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. *ArXiv*, abs/1805.11222, 2018.

Weihua Hu, R. Bansal, Kaidi Cao, Nikhil S. Rao, Karthik Subbian, and J. Leskovec. *Learning Backward Compatible Embeddings*. 2022.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings. *ArXiv*, abs/2505.12540, 2025.

Jeff Johnson, Matthijs Douze, and H. Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547, 2017.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, F. Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401, 2020.

Yury Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:824–836, 2016.

Lucas Maystre, Alvaro Ortega Gonzalez, Charles Park, Rares Dolga, Tudor Berariu, Yu Zhao, and Kamil Ciosek. When embedding models meet: Procrustes bounds and applications. *ArXiv*, abs/2510.13406, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084, 2019.

Yantao Shen, Yuanjun Xiong, W. Xia, and Stefano Soatto. Towards backward-compatible representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6367–6376, 2020.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663, 2021.

Harshil Vejendla. Drift-adapter: A practical approach to near zero-downtime embedding model upgrades in vector databases. *ArXiv*, abs/2509.23471, 2025.

Jinsung Yoon and Sercan Ö. Arik. Embedding-converter: A unified framework for cross-model embedding transformation. pp. 25464–25482, 2025.