The Medium Is Not the Message: Deconfounding Document Embeddings via Linear Concept Erasure

Anonymous ACL submission

Abstract

Embedding-based similarity metrics between text sequences can be influenced not just by the content dimensions we most care about, but can also be biased by spurious attributes like the text's source or language. These document confounders cause problems for many applications, but especially those that need to pool texts from different corpora. This paper shows that a debiasing algorithm that removes information about observed confounders from the encoder representations substantially reduces these biases at a minimal computational cost. Document similarity and clustering metrics improve across every embedding variant and task we evaluate—often dramatically. Interestingly, performance on out-of-distribution benchmarks is not impacted, indicating that the embeddings are not otherwise degraded.

1 Introduction

Suppose a political scientist is studying U.S. political discourse. They have access to data from two sources: Twitter posts from senators and summaries of congressional bills. A natural first step in data exploration is to first embed the texts (e.g., with a sentence transformer; Reimers and Gurevych 2019) and then cluster them (e.g., with k-means). As it turns out, some clusters will overwhelmingly contain items from one source or the other, because systematic differences between sources dominate the distances underpinning k-means (Fig. 1A).

Text embeddings are generated by pretrained models, and are able to capture topical, semantic, stylistic, multilingual, syntactic, and other information about the embedded text. Generally, models are trained with the goal of "making semantically similar sentences close in vector space" (Reimers and Gurevych, 2019). But pushing for this goal means that spurious correlations between attributes—e.g., domain and topic—can lead models to learn unwanted relationships. Per Thompson and Mimno (2018): "collections are often constructed by combining documents from multiple sources, [so the] most prominent patterns in a collection simply repeat the known structure of the corpus."[1] It would therefore seem useful to remove unwanted information from the representations.

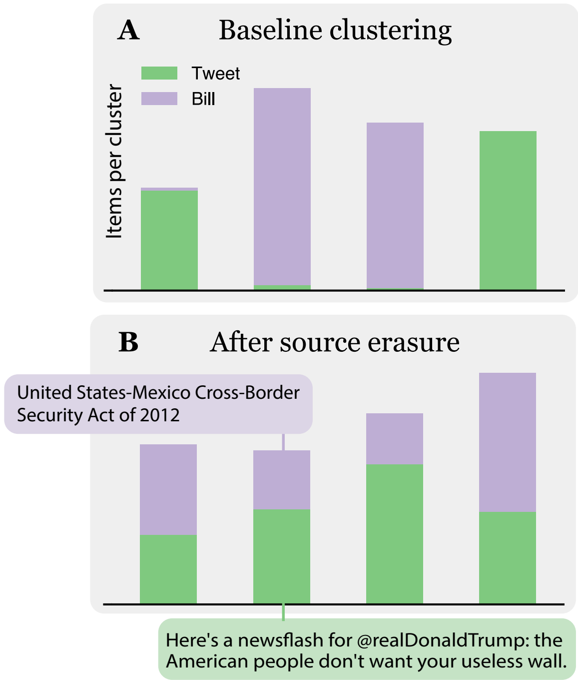

Figure 1: Clustering text embeddings from disparate sources (here, U.S. congressional bill summaries and senators' tweets) can produce clusters where one source dominates (Panel A). Using linear erasure to remove the source information produces more evenly balanced clusters that maintain semantic coherence (Panel B; sampled items relate to immigration). Four random clusters of k-means shown (k=25), trained on a combined 5,000 samples from each dataset.

Adjusting embeddings to remove confounding information is exactly what we do in this work.

[1] Their focus is on bag-of-words topic models rather than text embeddings, so their vocabulary-based approach does not translate to our setting.

Adapting the algorithm from Belrose et al. (2023) for linear concept erasure, we remove embedding subspaces that are predictive of the confounding variables that can bias measures of document distance. In the above example from U.S. politics, we residualize out the source information (Twitter or bills), producing adjusted embeddings for which similarity metrics load on the semantic content rather than the source (Fig. 1B). As another practical example, in a multilingual corpus, we residualize out the subspace that is predictive of language, leading to document distance metrics that are driven by content, rather than language.

We show through extensive tests that the adjusted embeddings perform significantly better for clustering and similarity search. For example, in a multilingual document search setting, Recall@1 increases from 0.175 to 0.826. Intriguingly, there is also no reduction in performance when using the adjusted embeddings on unseen datasets and tasks from a standard retrieval benchmark (Muennighoff et al., 2023; Enevoldsen et al., 2024), suggesting erasure does not harm embedding quality.

The approach is computationally inexpensive, involving only linear transformations on pretrained embeddings. Further, it can be used to adjust the embeddings for documents that don't have labels for the confounders. As a result, the method is particularly useful for applied work, for example in computational social science research. In sum, we:

- Formally show how erasure removes confounding information from document similarities (§2);
- Construct a benchmark of paired data designed to measure the effects of confounding attributes on embedding performance (§3);
- Evaluate a varied set of embedding methods, establishing that observable features, like a text's source, can harm the utility of text embeddings in applied settings (§4);
- Demonstrate that applying a linear erasure algorithm to remove observed confounders can mitigate such issues—sometimes dramatically—without impacting other aspects of performance (§5).[2]

2 Background

Many downstream tasks—nearest-neighbor search, clustering, retrieval, topic discovery—reduce to assessing how "close" two documents are in an embedding space. A good distance metric should rank pairs by semantic relatedness rather than by superficial attributes such as author, language, or publication venue. In practice, however, pretrained embedding models often encode these incidental signals because they appear frequently during training and help optimize their self-supervised objectives. When such signals correlate with content, distance measures become biased, undermining empirical conclusions drawn from them.

[2]We will release data and code upon acceptance.

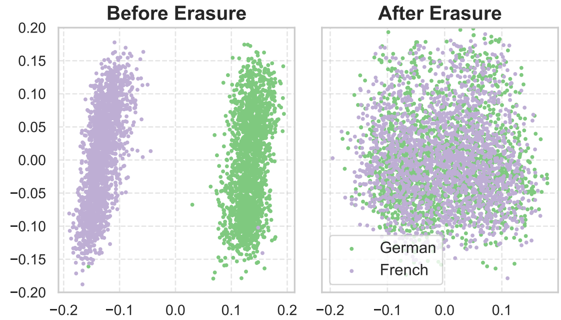

Figure 2: PCA projection of text embeddings before and after LEACE. Data are paired Swiss court case summaries in German (green) and French (purple). The first principal component recovers the two languages almost exactly.

Embedding text sequences Sentence-level embeddings place semantically close documents near one another in a vector space (Kiros et al., 2015; Conneau et al., 2017; Cer et al., 2018; Reimers and Gurevych, 2019). Modern systems start with a transformer encoder trained on masked-language modeling, then refine it on hundreds of millions of contrastive pairs drawn from many corpora (Reimers and Gurevych, 2019). This recipe powers state-of-the-art results in retrieval (Asai et al., 2021; Thakur et al., 2021; Zhang et al., 2023), clustering (Aggarwal and Zhai, 2012), and classification (Maas et al., 2011).

Contrastive batches often contain items from a single source so the model can focus on internal semantics (Nussbaum et al., 2024). A side-effect is that separate sources may occupy separate regions of the space, especially when cross-source positives are scarce. Multilingual models face a similar issue: even when trained with translation pairs (Wang et al., 2024), large amounts of monolingual data still push languages apart.

Notwithstanding efforts to make contrastive pairs comparable, the resulting embeddings still encode confounding information. Platform-specific jargon and style can be pivotal. Language can proxy for topic or geography. For authors or outlets, stylistic markers linked to gender or ideology become shortcuts for similarity. Because these attributes correlate with content, they act as *observed confounders* in distance-based analyses.

The document comparison problem. More formally, consider pairs of documents d_0 and d_1 with unit-norm embeddings $\boldsymbol{x_0}, \boldsymbol{x_1} \in \mathbb{R}^d$, $\|\boldsymbol{x_i}\| = 1$. Assume a linear decomposition for the embedding:

$$\boldsymbol{x}_i = \boldsymbol{B}_z \boldsymbol{z}_i + \boldsymbol{B}_c \boldsymbol{c}_i + \boldsymbol{B}_u \boldsymbol{u}_i + \varepsilon_i, \tag{1}$$

where $\boldsymbol{z}_i$ captures the *semantic content* of interest (e.g. topic); $\boldsymbol{c}_i$ collects *observed confounders* (source, language, author traits); $\boldsymbol{u}_i$ collects *unobserved confounders*; $\boldsymbol{B}_z$,$\boldsymbol{B}_c$, $\boldsymbol{B}_u$ are loading matrices; and ε_i is mean-zero noise uncorrelated with $(\boldsymbol{z}_i, \boldsymbol{c}_i, \boldsymbol{u}_i)$; we also assume these factors are zero-mean and have zero covariance with each other.

Similarity is measured with the dot product:

$$Y_{01} = Y(\boldsymbol{x}_0, \boldsymbol{x}_1) = \boldsymbol{x}_0^\top \boldsymbol{x}_1. \tag{2}$$

Taking expectations and using (1) gives

$$\mathbb{E}[Y_{01}] = \boldsymbol{z}_0^\top \boldsymbol{\Gamma}_z \boldsymbol{z}_1 + \boldsymbol{c}_0^\top \boldsymbol{\Gamma}_c \boldsymbol{c}_1 + \boldsymbol{u}_0^\top \boldsymbol{\Gamma}_u \boldsymbol{u}_1. \tag{3}$$

where $\boldsymbol{\Gamma}_k = \boldsymbol{B}_k^\top \boldsymbol{B}_k$. Only the first term reflects the semantic proximity we care about; the others bias any analysis based on Y_{01}.

Debiasing and concept erasure. Early debiasing work on word vectors identified a "bias direction" (e.g. gender) and removed its projection (Bolukbasi et al., 2016). Subsequent studies showed that the removed signal was still recoverable (Gonen and Goldberg, 2019), prompting stronger linear methods such as INLP (Ravfogel et al., 2020), LACE (Ravfogel et al., 2022), and LEACE (Belrose et al., 2023). These approaches search for an affine map that destroys all linear correlation with a protected attribute while moving points as little as possible.

An important special case of these kinds of concept erasure is *linear concept erasure*, where the goal is to prevent linear adversaries from predicting the information we aim to remove. This is usually achieved in the form of a projection matrix that neutralizes a subspace that is associated with the concept C. Following Ravfogel et al. (2022), Belrose et al. (2023) derived sufficient and necessary conditions for achieving *linear guardedness* (Ravfogel et al., 2023), a situation where *no linear classifier* can recover the concept C and achieve a loss lower than that of a trivial predictor that always predicts the majority class. Specifically, they derive a *linear projection matrix* P* such that:

$$\boldsymbol{P}* = \arg \min_{\mathbf{P} \in \mathbb{R}^{d \times d}} \mathbb{E}\left[||\boldsymbol{P}\mathbf{x} - \mathbf{x}||\right] \tag{4}$$

$$\text{subject to } \mathrm{Cov}(\boldsymbol{P}\mathbf{x}, C) = 0. \tag{5}$$

The covariance constraint ensures the erasure of linear information, while the first objective minimizes *distortion* of the representation space. It turns out that this objective has a closed-form solution in the form of

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{W}^\dagger (\boldsymbol{W}\boldsymbol{\Sigma}_{XC})(\boldsymbol{W}\boldsymbol{\Sigma}_{XC})^\dagger \boldsymbol{W} \tag{6}$$

where $\boldsymbol{W} = \boldsymbol{\Sigma}_{XX}^{-1/2}$ is a whitening matrix, and

$$\boldsymbol{\Sigma}_{XC} = \mathrm{Cov}(X, C), \boldsymbol{\Sigma}_{XX} = \mathrm{Cov}(X), \boldsymbol{\mu} = \mathbb{E}[X].$$

This condition is proved to be sufficient and necessary for achieving *linear guardedness*, i.e., the inability of any linear classifier to recover the attribute C from the embeddings.

In other words, Eq. (6) and $\boldsymbol{b} = \boldsymbol{\mu} - \boldsymbol{P}\boldsymbol{\mu}$ give the unique affine map that removes *all* linear correlation with the observed confounder C while altering the embeddings as little as possible.

For any (possibly unlabeled) document, the *adjusted* embedding $\tilde{\boldsymbol{x}}_i = P\boldsymbol{x}_i + \boldsymbol{b}$ has $\mathrm{Cov}(\tilde{\boldsymbol{x}}, C) = 0$ while minimizing the distance between $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$. Applying the LEACE map $\tilde{\boldsymbol{x}}_i = \boldsymbol{P}\boldsymbol{x}_i + \boldsymbol{b}$ to the structural decomposition in (1) gives

$$\begin{aligned} \tilde{X}_i &= \boldsymbol{P}\big(\boldsymbol{B}_Z \boldsymbol{z}_i + \boldsymbol{B}_C \boldsymbol{c}_i + \boldsymbol{B}_U \boldsymbol{u}_i + \varepsilon_i\big) + \boldsymbol{b} \\ &= \boldsymbol{B}_Z \tilde{\boldsymbol{z}}_i + \boldsymbol{B}_U \tilde{\boldsymbol{u}}_i \end{aligned} \tag{7}$$

where the middle term vanishes; since $\mathrm{Cov}(\boldsymbol{P}\boldsymbol{x}, C) = 0$, $\boldsymbol{B}_C = 0$. In turn, the estimand for the document similarity

$$\tilde{Y}_{01} = \tilde{\boldsymbol{x}}_0^\top \tilde{\boldsymbol{x}}_1 = \tilde{\boldsymbol{z}}_0^\top \boldsymbol{\Gamma}_{\boldsymbol{Z}} \tilde{\boldsymbol{z}}_1 + \tilde{\boldsymbol{u}}_0^\top \boldsymbol{\Gamma}_U \tilde{\boldsymbol{u}}_1, \tag{8}$$

is also purged of C. Note, however, that $\tilde{\boldsymbol{z}} = \boldsymbol{P}\boldsymbol{z}$ may not be equal to $\boldsymbol{z}$, depending on the intensity and nature of the dependence between $\boldsymbol{z}$ and C. So the LEACE algorithm might also add bias to similarity metrics through its adjustment of $\boldsymbol{z}$. Further, the (adjusted) unobserved confounder $\tilde{\boldsymbol{u}}$ remains, and it is unclear how the deconfounding by LEACE would either increase or reduce bias from $\boldsymbol{u}$.

3 Experimental Setup

Our evaluation settings are designed to approximate real-world use cases involving datasets from multiple corpora. They are divided into two groups, *category-level* and *event-level* data, both aiming to measure the same thing: the extent to which documents that share a common label have similar embeddings.

The approach is the same across all datasets: create a vector of concept labels C to erase, using known metadata (here, a text's source or language). Then, pass each text item through the embedding model to obtain a matrix $\mathbf{X}$. Fit LEACE on $(\mathbf{X}, C)$ to learn the whitening and projection matrices, then apply the transformations back to $\tilde{\mathbf{X}}$.[3]

3.1 Category-level Data

Category-level Data	N_{total}	Categories
CAP Data		
Bills – Orders	1,902	21
Bills – Newspapers	2,613	21
Orders – Newspapers	1,907	21
All Three Sources	3,211	21

Event-level Data	N_{paired}	N_{unpaired}
SCOTUS Cases		
Wikipedia – LexisNexis	2,048	1,518
Wikipedia – Oyez	1,560	1,762
LexisNexis – Oyez	2,048	2,075
SemEval News Articles		
EN – Non-EN	888	0
Swiss Court Cases		
DE – FR	2,048	1,760
DE – IT	2,048	1,760
FR – IT	2,048	1,760

Table 1: Dataset statistics. The data cover a variety of domains and languages.

Recalling the motivating example from the introduction, imagine a researcher clusters documents from different sources (like news articles and court cases), with the hope that each cluster contains documents that fall under a coherent topic.

We measure progress on this task by relying on a common set of ground-truth category labels, like "Education", that cover multiple datasets. The goal is that the assigned clusters align with the categories, even if the constituent documents come from different sources.

Datasets. We use datasets from the Comparative Agendas Project (CAP), which provides a coding framework for analyzing policy activities across time and between countries (Jones et al., 2023b).

We use texts from three sources: newspaper articles[4], congressional bill summaries (Wilkerson et al. 2023, taken from Hoyle et al. 2022), and executive orders (Jones et al., 2023a). We evaluate each pair of sources separately, as well as all three simultaneously.

Metrics and Methodology. We measure alignment between ground-truth category labels and assigned clusters with two metrics. Following Poursabzi-Sangdeh et al. (2016), we use purity, which quantifies to what extent each cluster contains items from a single gold category, and the Adjusted Rand Index, a chance-corrected metric that measures the similarity of two clusterings.

The erased concept is the *source* for each of the four settings (Table 1). When generating clusters, we follow a standard practice and apply k-means to the text embeddings for each document (Zhang et al., 2022).[5]

3.2 Event-level Data

Now imagine that a practitioner wants to understand how a common event—a court case, a natural disaster—is portrayed by distinct sources or languages. If they have access to one document discussing the event, how can they best find others?

Datasets. We rely on three paired datasets, which link documents depicting the same event in different sources or languages.

Super-SCOTUS (Fang et al., 2023) contains case summaries from the U.S. Supreme Court sourced from LexisNexis and Oyez. In addition, we scrape case summaries from Wikipedia. This results in 1,518 pairs of LexisNexis and Wikipedia case summaries, 2,075 from LexisNexis and Oyez, and 780 pairs from Wikpedia and Oyez.

SemEval 2022 Task 8 (Chen et al., 2022) assesses the similarity between pairs of multilingual news articles. We obtain 444 pairs of news articles that depict similar events in different languages, namely English and non-English (Spanish, German, and Chinese).

[3]For the out-of-sample experiments in Section 5, the transformations are applied to novel benchmark data $\mathbf{X}'$.

[4]https://comparativeagendas.net/project/pennsylvania

[5]We set $k = 21$, the total number of categories in the data. Improvements are robust to different k, see Fig. 9 in appendix.

Model	Bills & News				Orders & News				Bills & Orders				All Three Sources			
	Purity		ARI		Purity		ARI		Purity		ARI		Purity		ARI	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
MiniLM	0.346	**0.507**	0.148	**0.268**	0.329	**0.463**	0.123	**0.228**	0.391	**0.448**	0.169	**0.226**	0.269	**0.411**	0.096	**0.205**
GIST-small	0.380	**0.549**	0.171	**0.328**	0.421	**0.515**	0.200	**0.283**	0.422	**0.513**	0.191	**0.275**	0.330	**0.483**	0.131	**0.259**
E5-small	0.260	**0.414**	0.085	**0.207**	0.289	**0.290**	0.099	**0.101**	0.319	**0.422**	0.123	**0.190**	0.237	**0.356**	0.069	**0.166**
MPNet	0.365	**0.504**	0.162	**0.282**	0.377	**0.444**	0.151	**0.217**	0.461	**0.493**	0.229	**0.256**	0.334	**0.481**	0.130	**0.259**
GIST-base	0.373	**0.534**	0.157	**0.312**	0.380	**0.534**	0.165	**0.309**	0.425	**0.498**	0.188	**0.262**	0.320	**0.470**	0.054	**0.147**
E5-base	0.240	**0.375**	0.072	**0.175**	0.252	**0.297**	0.075	**0.108**	0.328	**0.407**	0.130	**0.173**	0.212	**0.346**	0.130	**0.173**
Nomic-v2	0.324	**0.463**	0.122	**0.250**	0.331	**0.353**	0.127	**0.161**	0.386	**0.442**	0.159	**0.218**	0.249	**0.411**	0.073	**0.196**
MXB-large	0.328	**0.493**	0.134	**0.279**	0.332	**0.524**	0.127	**0.281**	0.420	**0.487**	0.188	**0.263**	0.299	**0.410**	0.112	**0.199**
GIST-large	0.361	**0.492**	0.148	**0.295**	0.375	**0.471**	0.153	**0.258**	0.418	**0.495**	0.195	**0.258**	0.294	**0.434**	0.106	**0.226**
E5-large	0.224	**0.373**	0.066	**0.170**	0.273	**0.283**	0.082	**0.103**	0.327	**0.366**	0.104	**0.152**	0.211	**0.297**	0.055	**0.124**

Table 2: Cluster alignment metrics on the "category-level" Comparative Agendas Project datasets (§3.1), before and after linear concept erasure. Here, the erased concept is the *source* (top row). We set $k = 21$, the total number of categories in the CAP datasets. Erasure substantially improves cluster alignment for every combination of sources across all embedding models. Underlined scores indicate the highest value in each column.

A third dataset is derived from **SwilTra-Bench** (Niklaus et al., 2025), which contains parallel summaries of leading Swiss court decisions from the Federal Supreme Court of Switzerland in German, French, and Italian.

Methodology and Metrics. To accurately simulate real-world conditions, in which only partially paired data is available and the remaining data is unpaired and derived from different sources, we retain up to 1,024 data pairs for each applicable setting. We treat the remainder of the data as unpaired by randomly discarding one example from each pair. Thus, data is considered unpaired either because paired data was unavailable from the original sources or because one item from a pair was randomly removed. In each setting, we pool together the paired and unpaired data and subsequently use this combined dataset to train the LEACE eraser, aiming to remove source-specific information.

We evaluate whether each paired item can retrieve its counterpart from the pooled dataset using **Recall@1 and @10**, the proportion of correct matches that appear in the top k retrieved results.

3.3 Embedding Models

Our experiments use ten embedding models of varying sizes and dimensionality (appendix Table 12). This set includes multilingual and monolingual variants, as well as models with instruction fine-tuning: `MiniLM`[6], `GIST-small`, `GIST-base`, `GIST-large` (Solatorio, 2024), multilingual `E5-small`, `E5-base`, `E5-large` (Wang et al., 2024), `all-mpnet-base-v2` (Song et al., 2020), `Nomic-v2` (Nussbaum and Duderstadt, 2025), and `MXB-large` (Li and Li, 2023; Lee et al., 2024).

[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

4 Primary Results

We first discuss the results on the category-level datasets, then turn to the event-level. In brief, erasure improves embeddings across the board—over all models, metrics, and datasets we study.

4.1 Category-level

In all four source pairings from the CAP dataset, erasing source-specific information with LEACE consistently improves clustering quality (Table 2). In the *Bills–Newspapers* comparison, all ten models show marked improvements, with gains in ARI ranging from +0.104 (E5-large) to +0.157 (GIST-small), and purity increases as high as +0.169 (GIST-small). Although the magnitude of improvement varies, this pattern persists in the *Orders–Newspapers* comparison. While most models benefit substantially, multilingual models such as E5-small and E5-large show only marginal gains, suggesting that source signal may be less distinct in this pairing.

The *Bills–Orders* setting yields more moderate improvements, yet the gains remain consistent across model scales. Finally, the *All Three Sources* setting demonstrates that LEACE generalizes to more complex source distributions. Smaller-sized models, such as MiniLM and GIST-small, gain over +0.130 in purity and +0.100 in ARI. Even larger models such as GIST-large and MXB-large improve substantially after concept erasure.

Model	LexisNexis & Wikipedia				LexisNexis & Oyez				Oyez & Wikipedia			
	Recall@10		Recall@1		Recall@10		Recall@1		Recall@10		Recall@1	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
MiniLM	0.487	**0.606**	0.231	**0.313**	0.890	**0.899**	0.651	**0.693**	0.850	**0.924**	0.623	**0.747**
GIST-small	0.563	**0.656**	0.261	**0.325**	0.918	**0.943**	0.702	**0.778**	0.762	**0.844**	0.478	**0.599**
E5-small	0.421	**0.673**	0.176	**0.353**	0.830	**0.939**	0.563	**0.789**	0.689	**0.951**	0.398	**0.752**
MPNet	0.566	**0.666**	0.259	**0.337**	0.926	**0.943**	0.724	**0.775**	0.856	**0.911**	0.565	**0.678**
GIST-base	0.646	**0.757**	0.308	**0.412**	0.939	**0.963**	0.727	**0.819**	0.880	**0.950**	0.628	**0.773**
E5-base	0.414	**0.660**	0.188	**0.341**	0.830	**0.940**	0.575	**0.758**	0.650	**0.942**	0.371	**0.737**
Nomic-v2	0.530	**0.701**	0.254	**0.384**	0.950	**0.966**	0.770	**0.820**	0.903	**0.978**	0.658	**0.819**
MXB-large	0.537	**0.703**	0.249	**0.376**	0.928	**0.958**	0.720	**0.805**	0.883	**0.960**	0.654	**0.819**
GIST-large	0.657	**0.770**	0.305	**0.414**	0.954	**0.967**	0.787	**0.834**	0.947	**0.971**	0.760	**0.826**
E5-large	0.479	**0.720**	0.209	**0.381**	0.864	**0.949**	0.636	**0.791**	0.765	**0.964**	0.489	**0.792**

Table 3: Document similarity search results on paired "event-level" U.S. Supreme Court Summaries (3.2), before and after linear concept erasure. Here, the erased concept is the document's *source*. Erasure improves recall for every setting and model. Underlined scores indicate the highest value in each column.

Overall, these results demonstrate the robustness of LEACE across diverse source combinations and embedding models, confirming its ability to reduce spurious relationships between items while preserving task-relevant semantic structure.

4.2 Event-Level

At the event level, we present the results with Recall@10 and Recall@1, because only one document is deemed relevant for each query.

U.S. Supreme Court Case Summaries Applying LEACE consistently improves retrieval performance on the SCOTUS summary data (Table 10). In both *Wikipedia* pairings, improvements are large and especially pronounced for the E5 family. For instance, on *LexisNexis-Wikipedia*, E5-small gains +0.177 in Recall@1 and E5-base +0.153.

Performance before erasure on *LexisNexis–Oyez* is already high, likely because the two have more stylistic elements in common—both being technical summaries based on the original court opinion. Nonetheless, we still observe more modest but consistent gains. E5-small and E5-base increase Recall@1 by +0.226 and +0.183, respectively, although GIST-base and MXB-large exhibit improvements of only about +0.08.

Overall, LEACE not only improves representation consistency across heterogeneous legal sources, but also enhances alignment even when initial model performance is already strong.

Swiss Federal Supreme Court Case Summaries Turning now to multilingual data, we observe that LEACE can be extremely effective, even with already-multilingual embeddings (Table 4).

For all settings on the Swiss court case summary data, nearly every model sees higher recall after applying LEACE. The improvements tend to be largest with different language families: German-Italian and German-French. On *DE-IT*, gains in Recall@1 can reach +0.651 (E5-large); on *DE-FR*, +0.570 (E5-base). As French and Italian are closer, baseline retrieval is already strong, with some models already having near-perfect Recall@10. This reflects the tendency of related languages to lie closer in embedding space, as shown in prior work on genealogical structure (Östling and Kurfalı, 2023) and cross-lingual language representations (Sharoff, 2020). Still, increases in metrics abound, primarily in the smaller models like MiniLM. Taken together, LEACE removes source-specific signals even in complex multilingual legal domains.

SemEval News Articles To avoid bludgeoning the reader with positive results, we briefly outline the results on our other multilingual dataset: all ten models again benefit from erasure (Table 6 in the appendix).

5 Erasure helps, but can it hurt?

The results from the previous section appear conclusive: linear concept erasure removes spurious information from embeddings that distort similarities. At the same time, we might wonder: is it possible that erasure degrades the embeddings in subtle ways that our evaluations fail to detect? Although LEACE is designed to avoid unwanted distortions, it is possible the trained eraser removes "desirable" information that may support other tasks.

Model	DE & IT				DE & FR				FR & IT			
	Recall@10		Recall@1		Recall@10		Recall@1		Recall@10		Recall@1	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
MiniLM	0.009	**0.086**	0.003	**0.023**	0.026	**0.102**	0.008	**0.030**	0.146	**0.545**	0.040	**0.260**
GIST-small	0.020	**0.211**	0.004	**0.063**	0.041	**0.246**	0.011	**0.075**	0.315	**0.771**	0.101	**0.461**
E5-small	0.093	**0.930**	0.027	**0.543**	0.167	**0.937**	0.051	**0.563**	0.853	**1.000**	0.455	**0.968**
MPNet	0.016	**0.149**	0.006	**0.048**	0.053	**0.155**	0.021	**0.050**	0.157	**0.646**	0.052	**0.346**
GIST-base	0.034	**0.296**	0.008	**0.092**	0.076	**0.378**	0.024	**0.142**	0.440	**0.873**	0.167	**0.565**
E5-base	0.380	**0.987**	0.124	**0.749**	0.457	**0.989**	0.178	**0.748**	0.987	**1.000**	0.821	**0.979**
Nomic-v2	0.958	**0.994**	0.600	**0.765**	0.944	**0.996**	0.596	**0.767**	1.000	**1.000**	0.968	**0.979**
MXB-large	0.027	**0.356**	0.012	**0.117**	0.087	**0.427**	0.033	**0.168**	0.366	**0.910**	0.125	**0.632**
GIST-large	0.045	**0.298**	0.014	**0.090**	0.116	**0.385**	0.039	**0.152**	0.415	**0.880**	0.144	**0.551**
E5-large	0.503	**0.995**	0.175	**0.826**	0.722	**0.998**	0.300	**0.852**	0.988	**1.000**	0.831	**0.983**

Table 4: Document similarity search results on paired "event-level" multilingual Swiss Court Case Summaries (3.2), before and after linear concept erasure. Here, the concept is the document's *language*. Once again, erasure improves recall of the paired item in all cases, in some instances improving smaller models over their larger counterparts. Underlined scores indicate the highest value in each column.

In this section, we explore this question with additional evaluations on out-of-distribution (OOD) benchmarks. The experiments are designed to answer whether applying an eraser trained for a specific domain unintentionally harms general-purpose semantic representations. While our main experiments in the previous section target domain-specific differences, real-world deployment of embedding models often involves cross-domain tasks. We thus benchmark our models against diverse evaluation datasets from MTEB, (Muennighoff et al., 2023), determining whether erasers trained to isolate certain information also degrade performance in unrelated tasks.

5.1 Data and Methods

We focus on two sentence embedding models: MiniLM and E5-base-v2 (Wang et al., 2022). Each model is paired with two trained concept erasers: the CAP eraser, trained to remove the source from the *Bill–Newspapers* pair, and the Legal eraser, trained on *LexisNexis–Wikipedia*. This results in four models-eraser combinations per task.

We apply these combinations to retrieval and semantic textual similarity (STS) tasks from MTEB (Muennighoff et al., 2023): (1) Legal Retrieval tasks, (2) News Retrieval tasks (Thakur et al., 2021), and (3) STS News tasks. These benchmarks differ in domain, structure, and evaluation metrics, offering a comprehensive perspective on erased embedding behavior in out-of-domain settings. For each benchmark, we compare the performance of the original model embeddings to the same embeddings after applying the trained LEACE erasers.

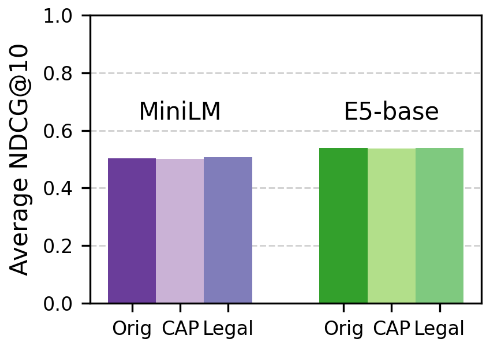

Figure 3: An eraser trained on embeddings from one dataset does not degrade embeddings from a different dataset. Erasers fit to the CAP and SCOTUS data (§3) are applied to embeddings (MiniLM and E5-base-v2) from five legal retrieval tasks. Each triplet of same-color bars compares the average NDCG@10 for the base and erased embeddings.

5.2 MTEB Results

We report a selection of results here, again emphasizing that our hope is not to improve benchmark results, but to *avoid making them worse* (full results in Appendix C).

Retrieval. On both the legal and news retrieval tasks, the trained erasers do not harm performance (as measured by the average NDCG@10). See Fig. 3 for legal retrieval; per-task performance (Fig. 5) and news retrieval (Fig. 6) are in the appendix. Given the domain overlap, we had hypothesized that the Legal eraser might improve

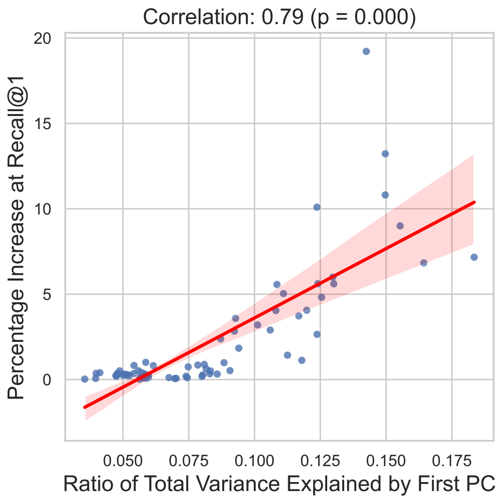

Figure 4: Relationship between variance explained by PC1 in the original embeddings and Recall@1 improvement after LEACE. Each point corresponds to a dataset setting in the event-level evaluation.

legal retrieval somewhat, but only one task sees a marginal improvement (AILACasedocs), from 0.197 to 0.218 (Table 7 in appendix). That said, the results are still positive overall, indicating that both the CAP and Legal erasers operate robustly in OOD retrieval tasks with both small and large models.

Semantic Textual Similarity. We evaluate the four model–eraser combinations on eight well-established STS benchmarks, covering both monolingual and crosslingual settings in the news domain (Fig. 7 and Table 9 in the appendix). The evaluation metric is the Spearman correlation between embedding cosine similarities and ground-truth semantic similarity. LEACE does not degrade performance over tasks, with most scores either unchanged or showing negligible increases.

Across the retrieval and semantic similarity evaluations, LEACE consistently preserves the quality of the embedding space while effectively removing targeted conceptual signals. These results reinforce its utility as a lightweight and reliable method for concept erasure.

6 Additional Findings

Relating LEACE to PCA Why does LEACE work in these settings? Here, we consider its relationship to Principal Component Analysis (PCA).

Taking the embeddings of the German-French Swiss court summaries, the first principal component (PC1) forms two clearly separable clusters corresponding directly to the text's language (Fig. 2). After applying LEACE, the clusters collapse into a single, overlapping distribution, an indication that language identity is no longer linearly separable in the embedding space.

To better understand when LEACE is effective, we investigate how the structural characteristics of the original embedding space relate to observed performance improvements. Specifically, we hypothesize that LEACE provides greater performance gains when the removable concept is prominently encoded within the embedding space.

We apply PCA to the original embeddings from each event-level dataset (SCOTUS, SemEval, Swiss Court Cases) and record the proportion of total variance explained by PC1. A high proportion of explained variance suggests that PC1 encodes a dominant direction in the embedding space, which will tend to correspond to the concept targeted by LEACE (i.e., the source or language, per Fig. 2).

Fig. 4 shows a strong positive correlation ($r = 0.79$, $p < 0.001$) between the proportion of variance explained by PC1 and the percentage improvement in Recall@1. This result indicates that LEACE is more effective when the removable concept aligns with dominant directions in the embedding space.

Given the above findings, why not use PCA to perform erasure instead, along the lines of Bolukbasi et al. (2016)? We find positive but less consistent results than LEACE on our tasks, and strongly degraded MTEB performance (Appendix E).

A new task: bitext mining Erasure improves already multilingual models on with multilingual tasks, so can it help with *bitext mining*—retrieving translation pairs via similarity search? Improvements are not as universally strong, but we do observe state-of-the-art results on a few leaderboard tasks from Enevoldsen et al. (2025), and erasure never harms performance (details in Appendix A).

7 Conclusion

For applied practitioners working with large text collections from multiple sources or languages—a regular occurrence—our results tell a fairly unambiguous story: apply linear erasure to any document embeddings before working with them to remove confounding information. While there are cases where it may not work, it does not seem to damage representations (see below), and comes at a minimal computational cost.

8 Limitations

The primary limitation of our method is its dependence on per-document metadata or labels. If some undesirable low-level pattern in the data distribution is suspected but not known—say, an unreported change in how a corpus was collected over a long time period—then a user will have to first apply some possibly-unsupervised labeling method.

Another shortcoming arises when metadata is available but the categories are too numerous relative to the total number of items. For instance, the paired *within*-language (en–en) SemEval Task 8 news articles come from dozens of sources, with many sources only being represented by a handful of articles. In contrast to removing the language in the multilingual data (Table 6), removing the source label does not improve retrieval results over the baseline. A possible direction for future work is to first combine similar sources into larger categories (e.g., local vs. national newspapers), then erase the category label.

A final limitation was first noted by Huang et al. (2024). They use LEACE as a baseline in multilingual retrieval contexts, removing language information as we do, but find mixed results. Hence, LEACE may not always help in all contexts. One initial hypothesis is that our tasks, while realistic, differ from the standard benchmark data that models are trained on, which may lead to saturated in-domain performance that does not transfer out-of-domain. It may also be that retrieval setups, with the distinct (short query, document) rather than our (document, document) structure, have characteristics that make them less amenable to erasure. In future work, we plan to explore these hypotheses and find an explanation for such discrepancies.

References

Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. *Mining text data*, pages 77–128.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2020. Aila 2019 precedent & statute retrieval task.

Ergun Biçici. 2015. RTM-DCU: Predicting semantic similarity with referential translation machines. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 56–63, Denver, Colorado. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and A. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural Information Processing Systems*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. Climate-fever: A dataset for verification of real-world climate claims. *Preprint*, arXiv:2012.00614.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Kenneth Enevoldsen, Emil Trenckner Jessen, and Rebekah Baglini. 2024. DANSK: Domain generalization of Danish named entity recognition. *Northern European Journal of Language Technology*, 10:14–29.

Biaoyan Fang, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2023. Super-SCOTUS: A multi-sourced dataset for the Supreme Court of the US. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 202–214, Singapore. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Nils Holzenberger and Benjamin Van Durme. 2021. Factoring statutory reasoning as language understanding challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2742–2758, Online. Association for Computational Linguistics.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiqi Huang, Puxuan Yu, Shauli Ravfogel, and James Allan. 2024. Language concept erasure for language-invariant dense retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13261–13273, Miami, Florida, USA. Association for Computational Linguistics.

Bryan D. Jones, Frank R. Baumgartner, Sean M. Theriault, Derek A. Epp, Rebecca Eissler, Cheyenne Lee, and Miranda E. Sullivan. 2023a. Policy agendas project: Executive orders.

Bryan D Jones, Frank R Baumgartner, Sean M Theriault, Derek A Epp, Cheyenne Lee, Miranda E Sullivan, and Chris Cassella. 2023b. Policy agendas project: codebook. *Comparative Agendas Project. URL: https://www. comparativeagendas. net/us*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread-new fluffy embeddings model.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27:117–139.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Laura Manor and Junyi Jessy Li. 2019. Plain English summarization of contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Philip May. 2021. Machine translated multilingual sts benchmark dataset.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Joel Niklaus, Jakob Merane, Luka Nenadic, Sina Ahmadi, Yingqiang Gao, Cyrill AH Chevalley, Claude Humbel, Christophe Gösken, Lorenzo Tanzi, Thomas Lüthi, et al. 2025. Swiltra-bench: The swiss legal translation benchmark. *arXiv preprint arXiv:2503.01372*.

Zach Nussbaum and Brandon Duderstadt. 2025. Training sparse mixture of experts text embedding models. *Preprint*, arXiv:2502.07972.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Robert Östling and Murathan Kurfalı. 2023. Language embeddings sometimes contain typological generalizations. *Computational Linguistics*, 49(4):1003–1051.

Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. 2016. ALTO: Active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1158–1169, Berlin, Germany. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. 2023. Log-linear guardedness and its implications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9413–9431, Toronto, Canada. Association for Computational Linguistics.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022. Linear adversarial concept erasure. In *International Conference on Machine Learning*.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. 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 4947–4958, Hong Kong, China. Association for Computational Linguistics.

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.

Serge Sharoff. 2020. Finding next of kin: Cross-lingual embedding spaces for related languages. *Natural Language Engineering*, 26(2):163–182.

Aivin V Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Laure Thompson and David Mimno. 2018. Authorless topic models: Biasing models away from known structure. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3903–3914, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Steven Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dmitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. MAUD: An expert-annotated legal NLP dataset for merger agreement understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16369–16382, Singapore. Association for Computational Linguistics.

Steven H. Wang, Maksim Zubkov, Kexin Fan, Sarah Harrell, Yuyang Sun, Wei Chen, Andreas Plesner, and Roger Wattenhofer. 2025. Acord: An expert-annotated retrieval dataset for legal contract drafting. *Preprint*, arXiv:2501.06582.

Orion Weller, Benjamin Chang, Eugene Yang, Mahsa Yarmohammadi, Samuel Barham, Sean MacAvaney, Arman Cohan, Luca Soldaini, Benjamin Van Durme, and Dawn Lawrie. 2025. mfollowir: A multilingual benchmark for instruction following in retrieval. In *European Conference on Information Retrieval*, pages 295–310. Springer.

John Wilkerson, E Scott Adler, Bryan D Jones, Frank R Baumgartner, Guy Freedman, Sean M Theriault, Alison Craig, Derek A Epp, Cheyenne Lee, and Miranda E Sullivan. 2023. Policy agendas project: Congressional bills.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, Berlin, Germany. Association for Computational Linguistics.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66.

A Bitext Mining Results

Our gains on multilingual tasks with already multilingual embeddings motivate us to ask whether erasure can benefit already "saturated" leaderboard tasks that cover multiple languages. To this end, we focus on *bitext mining*: given pairs of sentences in different languages, the goal is to retrieve a specific sentence in the target language given a "query" sentence in the source language (typically a translation; F_1 is the standard metric). We collect all 28 tasks available through the MTEB package at the time of writing (Muennighoff et al., 2022) and use `E5-large-instruct`, one of the best-performing models on the leaderboard.

In several cases, there is a marked increase, yielding state-of-the-art scores on three tasks that appear on the public leaderboard (even with a different base model class, Table 5 in appendix). Generally, though, the improvements are much smaller than those in our main experiments, with over half of the 28 tasks showing less than a 0.01 change (although no tasks decrease more than -0.01). First applying LEACE is therefore a simple step when bitext mining; even if it may not always help, it is unlikely to hurt.

	F_1		
	Before	After	Δ
SynPerChatbotSumS	0.283	0.500	0.217
SAMSumFa	0.811	0.943	0.132
SynPerChatbotRAGSumS	0.560	0.680	0.120
RomaTales	0.201	0.263	0.062
SRNCorpus	0.500	0.551	0.051
NusaX*	0.853	**0.892**	0.039
NollySenti*	0.807	**0.839**	0.032
NusaTranslation*	0.851	**0.876**	0.025
LinceMT	0.487	0.506	0.019
Bornholm*	0.560	0.578	0.018
IN22Conv	0.626	0.637	0.011
Phinc	0.855	0.867	0.011
Number of tasks with $\lvert\Delta\rvert < 0.01$			*15*

Table 5: F_1 on MTEB Bitext Mining Tasks before and after erasing the language ID, for `E5-large-instruct`. Gains are substantial in a few cases, sometimes improving over the reported state-of-the-art on MTEB (tasks with * appear on the public leaderboard, improvements over SotA in **bold**).

B SemEval English & Non-English News Results

The results on testing LEACE on the SemEval 2022 Task 8 dataset are presented in Table 6. All models benefit from LEACE, with consistent improvements in both Recall@10 and Recall@1. The E5-small model shows the strongest gains overall: +0.202 (Recall@10) and +0.236 (Recall@1). High-performing large models like E5-large and MXB-large achieve further enhancements of up to +0.156 in Recall@1. Smaller models also gain notable increases. For instance, MiniLM gains +0.183 (Recall@10) and +0.127 (Recall@1), respectively. These improvements highlight LEACE's utility in reducing source bias and improving semantic alignment in multilingual event representations. Nomic-v2, which already has high scores before LEACE, showed modest increases, likely due to saturation. In general, LEACE proves effective even under high-resource, multilingual scenarios.

	`Recall@10`		`Recall@1`	
Model	Before	After	Before	After
MiniLM	0.350	**0.533**	0.150	**0.277**
GIST-small	0.497	**0.636**	0.247	**0.372**
E5-small	0.614	**0.816**	0.318	**0.554**
MPNet	0.557	**0.664**	0.262	**0.347**
GIST-base	0.564	**0.694**	0.301	**0.402**
E5-base	0.777	**0.859**	0.466	**0.601**
Nomic-v2	0.892	**0.906**	0.637	**0.651**
MXB-large	0.527	**0.691**	0.250	**0.390**
GIST-large	0.624	**0.734**	0.332	**0.428**
E5-large	0.747	**0.866**	0.436	**0.592**

Table 6: Results on SemEval English & Non-English News Articles

C MTEB Evaluation Results

We report the full evaluation results of the CAP and Legal erasers on three MTEB benchmark groups: Legal Retrieval, News Retrieval, and STS News Tasks. Each setting involves comparing model performance before and after LEACE-based erasure, across two embedding models (MiniLM and E5-base), as shown in Table 7, Table 8, and Table 9 and Fig. 5, Fig. 6 and Fig. 7.

D Sources of MTEB Tasks

We list below the original sources for the datasets used from the MTEB benchmark (Muennighoff et al., 2023; Enevoldsen et al., 2025):

- Legal retrieval tasks: **AILACasedocs** and **AILAStatutes** (Bhattacharya et al., 2020), **LegalBenchConsumerContractsQA** (Wang et al., 2025; Koreeda and Manning, 2021), **LegalBenchCorporateLobbying** (Guha et al., 2023; Holzenberger and Van Durme, 2021; Lippi et al., 2019; Ravichander et al., 2019; Wang et al., 2023; Wilson et al., 2016;

Task	MiniLM			E5-base		
	Before	After (CAP)	After (Legal)	Before	After (CAP)	After (Legal)
`AILACasedocs`	0.197	0.197	0.218	0.292	0.290	0.292
`AILAStatutes`	0.205	0.196	0.205	0.186	0.191	0.193
`ConsumerContractsQA`	0.656	0.659	0.654	0.720	0.712	0.720
`CorporateLobbying`	0.864	0.865	0.863	0.915	0.914	0.913
`LegalSummarization`	0.590	0.591	0.592	0.577	0.576	0.578

Table 7: Legal Retrieval Results on MTEB evaluated using NDCG@10. Each model (MiniLM, E5-base-v2) is tested with and without LEACE erasure, using both CAP and Legal erasers.

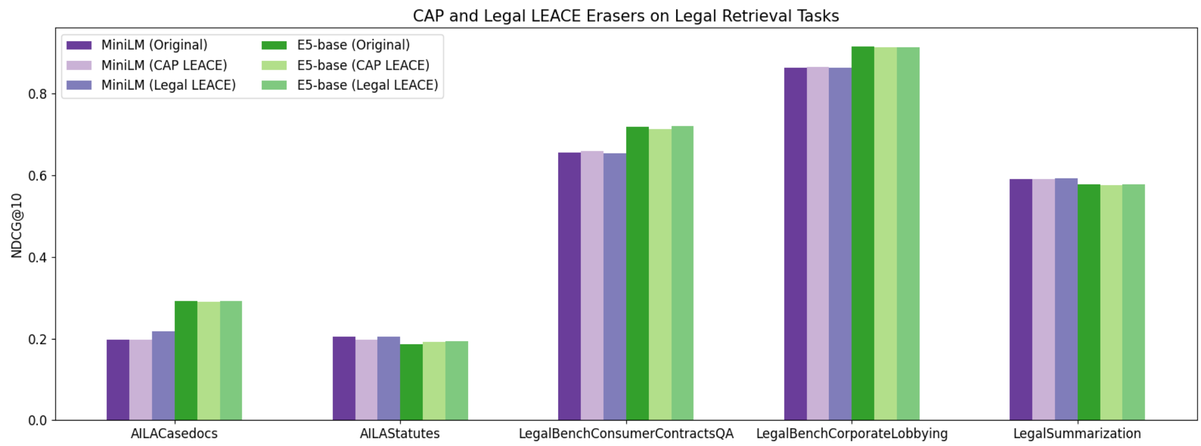

Figure 5: Performance of CAP and Legal erasers across three news retrieval tasks. Each group of bars compares the base and LEACE-erased models for MiniLM and E5-base-v2 embeddings.

Zheng et al., 2021; Zimmeck et al., 2019), **LegalSummarization** (Manor and Li, 2019).

- News retrieval tasks: **BelebeleRetrieval** (Bandarkar et al., 2024), **NanoClimate-FeverRetrieval** (Diggelmann et al., 2021), **mFollowIRCrossLingualInstructionRe-trieval** (Weller et al., 2025).

- STS news tasks: **IndicCrosslingualSTS** (Ramesh et al., 2022), **STS12** (Agirre et al., 2012), **STS13** (Agirre et al., 2013), **STS15** (Biçici, 2015), **STS17** (Cer et al., 2017), **STS22** (Chen et al., 2022), **STSBenchmark** and **STSBenchmarkMultilingualSTS** (May, 2021).

E Additional PCA Analysis

We create a baseline by removing PC1 from the embedding space, and evaluate it in the event-level setting using the SCOTUS dataset (Table 10). Overall, the baseline occasionally helps and can even marginally outperform LEACE in a few cases, but its effectiveness appears unstable, heavily dependent on the particular setting and model used (although it is effective for the E5 family for most configurations). Furthermore, in some cases, it performs worse than applying no erasure at all. There is also a final catch: removing the learned PC1 from OOD embeddings *does* dramatically degrade performance on MTEB tasks (Table 11), unlike LEACE (Fig. 8).

E.1 Event-Level Results on SCOTUS Case Summaries

Table 10 reveals the results of applying the baseline, which removes the first principal component (PC1) from the embedding space, in the event-level setting on the SCOTUS dataset. While it sometimes improves over the original embeddings and occasionally outperforms LEACE (especially for the E5 family), its performance is inconsistent across models and configurations, and it can underperform even relative to no erasure.

E.2 MTEB Evaluation Results

Table 11 shows the results of applying the baselines, derived from both CAP and SCOTUS datasets, on

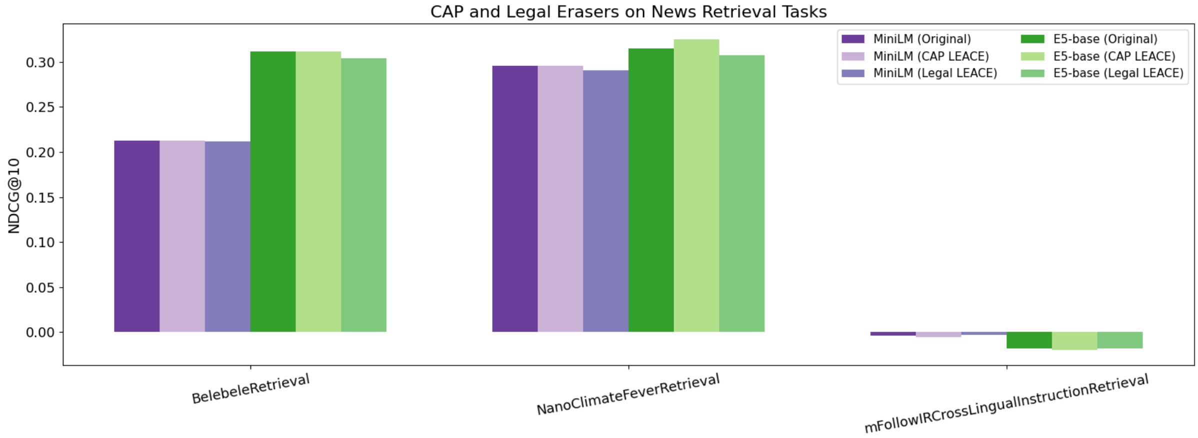

Figure 6: Performance of CAP and Legal erasers across three news retrieval tasks. Each group of bars compares the base and LEACE-erased models for MiniLM and E5-base-v2 embeddings.

Task	MiniLM			E5-base		
	Before	After (CAP)	After (Legal)	Before	After (CAP)	After (Legal)
BelebeleRetrieval	0.212	0.212	0.211	0.312	0.311	0.303
NanoClimateFeverRetrieval	0.296	0.296	0.291	0.315	0.325	0.307
mFollowIR (CrossLingual)	-0.004	-0.005	-0.003	-0.018	-0.019	-0.018

Table 8: News Retrieval Results on MTEB evaluated using NDCG@10. Each model (MiniLM, E5-base-v2) is evaluated before and after applying LEACE, using both CAP and Legal erasers.

the MTEB legal retrieval tasks. In all cases, this PC1 removal leads to a drastic performance drop for both MiniLM and E5-base models. As observed in the comparison between the two approaches in Fig. 8, in contrast, LEACE erasures maintain retrieval quality, highlighting its robustness.

F Embedding model information

We list characteristics of the embedding models in Table 12.

G Use of AI Assistants

We used AI assistants, including ChatGPT and Claude, for editing (e.g., grammar, spelling, word choice), debugging code, and visualizing results for submission.

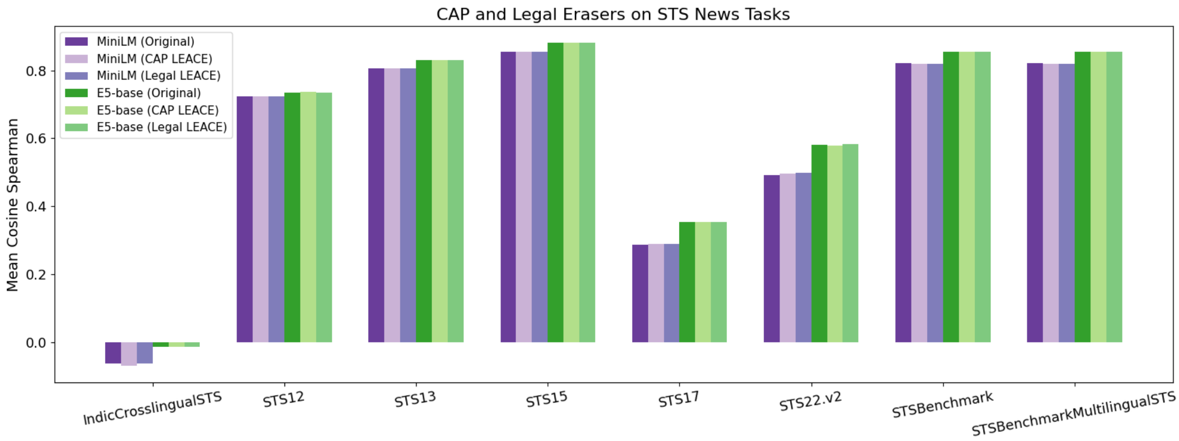

Figure 7: Performance of CAP and Legal erasers across eight STS news tasks. Each group of bars compares the base and LEACE-erased models for MiniLM and E5-base-v2 embeddings.

Task	MiniLM			E5-base		
	Before	After (CAP)	After (Legal)	Before	After (CAP)	After (Legal)
`IndicCrosslingualSTS`	-0.063	-0.070	-0.062	-0.013	-0.012	-0.013
`STS12`	0.724	0.724	0.723	0.735	0.736	0.735
`STS13`	0.806	0.806	0.806	0.830	0.830	0.830
`STS15`	0.854	0.854	0.854	0.882	0.882	0.882
`STS17`	0.288	0.289	0.288	0.354	0.355	0.353
`STS22.v2`	0.492	0.496	0.499	0.581	0.578	0.583
`STSBenchmark`	0.820	0.820	0.820	0.855	0.855	0.855
`STSBenchmarkMultilingualSTS`	0.820	0.820	0.820	0.855	0.855	0.855

Table 9: STS News Results on MTEB evaluated using the mean cosine Spearman score. Each model (MiniLM, E5-base-v2) is evaluated before and after LEACE, using both CAP and Legal erasers.

Model	LexisNexis & Wikipedia						LexisNexis & Oyez						Oyez & Wikipedia					
	`Recall@10`			`Recall@1`			`Recall@10`			`Recall@1`			`Recall@10`			`Recall@1`		
	Before	After	Baseline	Before	After	Baseline	Before	After	Baseline	Before	After	Baseline	Before	After	Baseline	Before	After	Baseline
MiniLM	0.487	**0.606**	0.478	0.231	**0.313**	0.231	0.890	**0.899**	0.883	0.651	**0.693**	0.646	0.850	**0.924**	0.869	0.623	**0.747**	0.670
GIST-small	0.563	**0.656**	0.547	0.261	**0.325**	0.254	0.918	**0.943**	0.920	0.702	**0.778**	0.701	0.762	**0.844**	0.776	0.478	**0.599**	0.500
E5-small	0.421	0.673	**0.675**	0.176	0.353	**0.356**	0.830	**0.939**	**0.939**	0.563	**0.789**	**0.789**	0.689	**0.951**	0.950	0.398	0.752	**0.753**
MPNet	0.566	**0.666**	0.552	0.259	**0.337**	0.257	0.926	**0.943**	0.925	0.724	**0.775**	0.722	0.856	**0.911**	0.862	0.565	**0.678**	0.574
GIST-base	0.646	**0.757**	0.636	0.308	**0.412**	0.309	0.939	**0.963**	0.936	0.727	**0.819**	0.725	0.880	**0.950**	0.917	0.628	**0.773**	0.701
E5-base	0.414	**0.660**	**0.660**	0.188	0.341	**0.344**	0.830	**0.940**	0.939	0.575	**0.758**	0.755	0.650	**0.942**	**0.942**	0.371	0.737	**0.738**
Nomic-v2	0.530	0.701	**0.703**	0.254	**0.384**	0.382	0.950	**0.966**	0.948	0.770	**0.820**	0.767	0.903	0.978	**0.981**	0.658	**0.819**	0.818
MXB-large	0.537	**0.703**	0.627	0.249	**0.376**	0.321	0.928	**0.958**	0.933	0.720	**0.805**	0.729	0.883	**0.960**	0.919	0.654	**0.819**	0.737
GIST-large	0.657	**0.770**	0.641	0.305	**0.414**	0.300	0.954	**0.967**	0.954	0.787	**0.834**	0.787	0.947	**0.971**	0.944	0.760	**0.826**	0.772
E5-large	0.479	**0.720**	0.717	0.209	0.381	**0.388**	0.864	**0.949**	**0.949**	0.636	**0.791**	0.790	0.765	**0.964**	0.963	0.489	**0.792**	**0.792**

Table 10: Event-Level Results on SCOTUS Case Summaries

Task	MiniLM			E5-base		
	Before	After (CAP)	After (Legal)	Before	After (CAP)	After (Legal)
`AILACasedocs`	0.197	0.039	0.044	0.292	0.027	0.042
`AILAStatutes`	0.205	0.082	0.092	0.186	0.081	0.079
`ContractsQA`	0.656	0.018	0.029	0.720	0.022	0.028
`CorporateLobbying`	0.864	0.012	0.016	0.915	0.012	0.004
`LegalSummarization`	0.590	0.011	0.012	0.577	0.018	0.006

Table 11: Legal Retrieval Results on MTEB evaluated using NDCG@10. Each mode (MiniLM, E5-base-v2) is evaluated before and after applying baseline model (PC1 removal), using both CAP and Legal erasers.

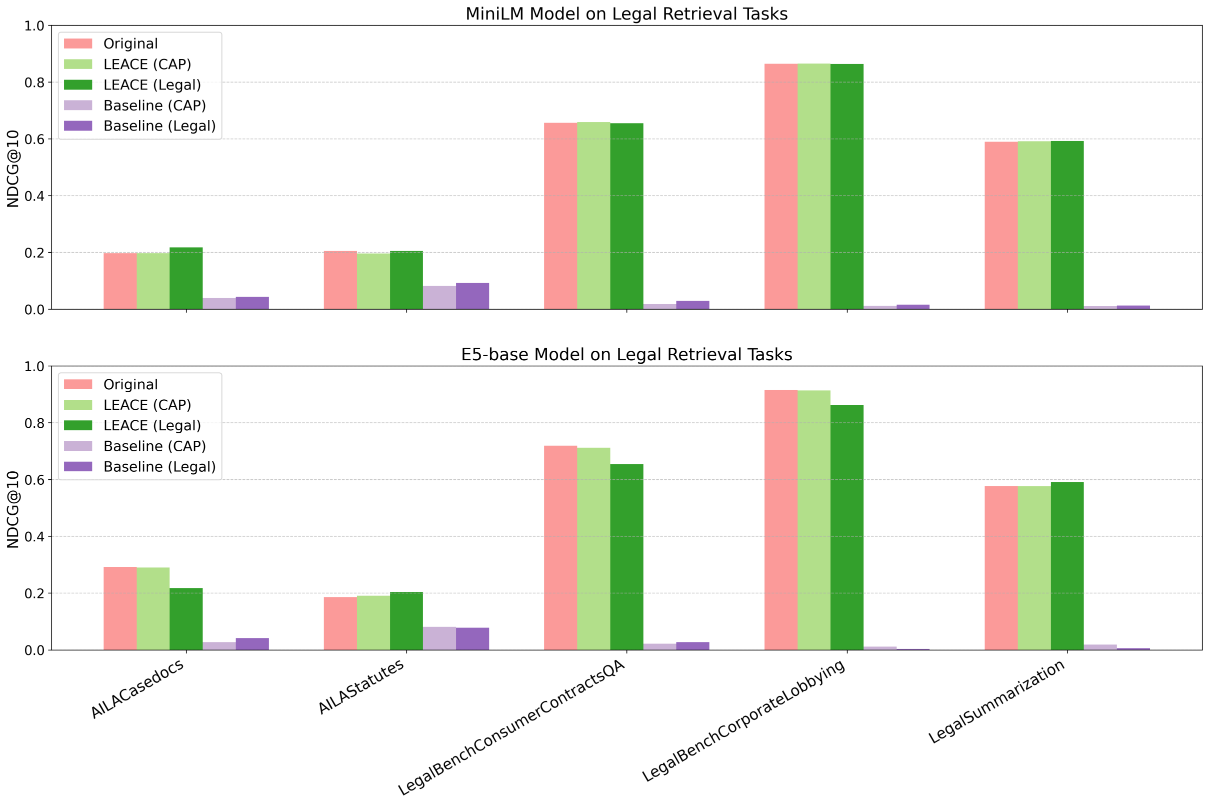

Figure 8: Comparison of average NDCG@10 scores across five MTEB legal retrieval tasks. Each group of bars compares the original, LEACE-erased and baseline models for MiniLM and E5-base-v2 models.

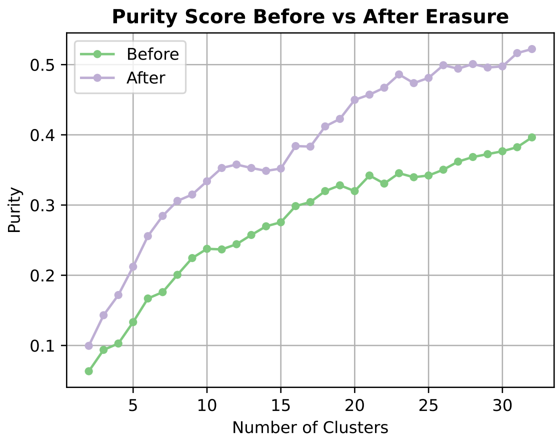

Figure 9: Purity score before vs. after LEACE erasure under different cluster counts, using data from CAP news articles and congressional bills.

Models	#Dims	#Params	Multilingual	IFT
MiniLM	384	22.7M		
GIST-small	384	33.4M		
E5-small	384	118M	✓	
MPNet	768	109M		
GIST-base	768	109M		
E5-base	768	278M	✓	
Nomic-v2	768	475M	✓	✓
MXB-large	1,024	335M		✓
GIST-large	1,024	335M		
E5-large	1,024	560M	✓	

Table 12: Embedding Models. We examine mono- and multilingual models spanning multiple parameter sizes and embedding dimensions.