LANGSAE EDITING: Improving Multilingual Information Retrieval via Post-hoc Language Identity Removal

Anonymous ACL submission

Abstract

Dense retrieval in multilingual settings often searches over mixed-language collections, yet multilingual embeddings encode language identity alongside semantics. This language signal can inflate similarity for same-language pairs and crowd out relevant evidence written in other languages. We propose LANGSAE EDITING, a post-hoc sparse autoencoder trained on pooled embeddings that enables controllable removal of language-identity signal directly in vector space. The method identifies language-associated latent units using cross-language activation statistics, suppresses these units at inference time, and reconstructs embeddings in the original dimensionality, making it compatible with existing vector databases without retraining the base encoder or re-encoding raw text. Experiments across multiple languages show consistent improvements in ranking quality and cross-language coverage, with especially strong gains for script-distinct languages. The LANGSAE model and training code are publicly available.[1]

1 Introduction

Dense retrieval ranks documents by comparing query and document embeddings, typically with cosine similarity, and it is a core component of modern search and retrieval-augmented generation pipelines (Karpukhin et al., 2020; Xiong et al., 2021; Khattab and Zaharia, 2020). In multilingual deployments, the indexed collection is often mixed-language, and relevant evidence for a query can appear in any language. In this setting, dense retrievers commonly exhibit a *same-language preference*, where same-language candidates receive a similarity advantage and crowd out more relevant evidence written in other languages (Yang et al., 2021, 2025).

We study this setting as multilingual information retrieval (MLIR): queries may be issued in any supported language and retrieval is performed against a single multilingual pool (Zhang et al., 2023, 2021). The failure mode is a mismatch with the goal of embedding-based retrieval, which is to prioritize semantic alignment rather than language match.

Prior analyses point to a concrete mechanism. Multilingual encoders encode language identity in addition to semantics, and language identity remains recoverable from their representations (Devlin et al., 2019; Conneau et al., 2020; Libovick'y et al., 2020, 2019). When similarity search is performed in a shared space, this language signal can distort neighborhood structure by inflating same-language cosine similarity, producing **Language Identity Bias in MLIR** (Yang et al., 2021, 2025).

Mitigating this bias is constrained by deployment reality. Many systems rely on precomputed document embeddings in a vector database, and encoder-side mitigation typically requires fine-tuning and then re-encoding the entire corpus from raw text, which is often the dominant cost in real deployments. This motivates post-hoc methods that operate directly on existing vectors and remain compatible with standard similarity search infrastructure (Johnson et al., 2019; Malkov and Yashunin, 2020).

We propose LANGSAE, a post-hoc method that suppresses language-identity signal in pooled embeddings while preserving retrieval-relevant semantics. LANGSAE is an overcomplete sparse autoencoder trained on pooled embeddings, its sparse feature representation enables language-associated factors to concentrate into a small set of latent units that can be selectively suppressed, then decoded back to the original embedding dimensionality for drop-in cosine scoring. Because the transformation is vector-only, it is substantially cheaper than encoder tuning and corpus-wide re-encoding, editing an embedding in 0.0445 ms, enabling both offline

[1] To preserve anonymity during review, we will provide the links after acceptance.

retrofitting of stored vectors and query-time editing.

Across Belebele (Bandarkar et al., 2024) and XQuAD (Artetxe et al., 2020), LANGSAE EDITING improves macro-average nDCG@20 by about +21.9% and +20.6%, respectively. Gains are especially large for script-distinct languages such as Chinese, consistent with language identity acting as a similarity shortcut in multilingual pools.

We make three contributions:

- We formalize **Language Identity Bias in MLIR** as same-language crowding in shared multilingual pools and introduce diagnostics that isolate this effect beyond aggregate retrieval metrics.
- We introduce LANGSAE EDITING, a sparse feature-based post-hoc transformation that suppresses language-associated units and reconstructs embeddings in the original space for drop-in retrieval.
- We demonstrate consistent gains on multilingual pools across two benchmarks and provide analyses that connect feature suppression to improved ranking behavior, with a lightweight transformation that is practical for retrofitting existing vector databases.

2 Related Work

2.1 Multilingual Dense Retrieval

Dense retrieval embeds queries and documents into a shared space and ranks by vector similarity (Karpukhin et al., 2020). Multilingual retrievers typically build on pretrained multilingual encoders (Devlin et al., 2019; Conneau et al., 2020) and are evaluated on multilingual retrieval datasets such as mMARCO, Mr. TyDi, and MIRACL (Bonifacio et al., 2021; Zhang et al., 2021, 2023), with broad embedding evaluations increasingly standardized by MTEB (Muennighoff et al., 2022). Recent improvements come from multilingual sentence embedding alignment (Artetxe and Schwenk, 2019; Feng et al., 2022), weakly supervised contrastive pretraining (Wang et al., 2022, 2024), unsupervised pretraining for multilingual dense retrieval (Wu et al., 2022), lightweight inference-time adaptation (Huang et al., 2023), contrastive objectives for language-agnostic retrieval (Hu et al., 2023), and distillation-based transfer (Yang et al., 2024). Despite these advances, retrieval quality often varies substantially across languages, motivating analyses of language-linked failure modes in multilingual IR (Yang et al., 2025).

2.2 Language Signal and Bias in Multilingual Representations

Language identity remains recoverable from multilingual representations, indicating that embeddings mix semantics with language-correlated structure (Libovick'y et al., 2019, 2020). Related work studies when cross-lingual transfer emerges and how representation spaces align across languages (Artetxe et al., 2020; Artetxe and Schwenk, 2019), and proposes reducing self-language preference by explicitly removing language information (Yang et al., 2021). Other approaches encourage language-agnostic structure during training (Zhao et al., 2021) or identify language-associated subspaces that can be filtered (Xie et al., 2022), while recent dense retrieval work explores language-invariant behavior through language concept erasure (Huang et al., 2024). From an IR perspective, language bias is also framed as an evaluation and fairness issue, where per-language reporting and disparity-aware analysis are important (Bandarkar et al., 2024; Yang et al., 2025).

2.3 Post-hoc Representation Transformation

Post-processing methods can improve cosine-neighborhood geometry by addressing anisotropy or dominant directions in embedding spaces, often via simple transformations such as whitening (Li et al., 2020; Huang et al., 2021). Autoencoder-based objectives offer an alternative, learning a transformation that reconstructs vectors while enabling controlled edits in latent space, as demonstrated by reconstruction-based sentence embedding learning (Wang et al., 2021). Our work follows this post-hoc direction but targets a specific nuisance factor, language identity, via sparse overcomplete features that can be selectively suppressed while reconstructing embeddings back to the original dimensionality for drop-in retrieval use.

3 Methodology

We propose LANGSAE, a post-hoc method that edits pooled encoder embeddings to suppress language-identity signal while preserving retrieval-relevant semantics. We assume a standard dense retrieval pipeline where a frozen multilingual encoder produces token representations, which are

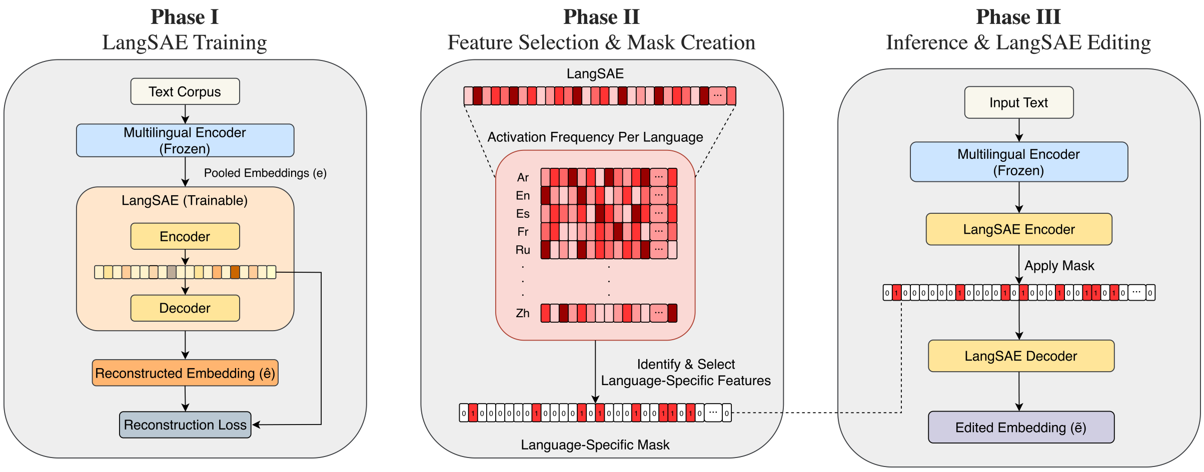

Figure 1: Overview of the LANGSAE EDITING pipeline. **Phase I:** Train an overcomplete sparse autoencoder on pooled embeddings from a frozen multilingual encoder. **Phase II:** Compute per-language activation frequencies and select language-associated features to form a mask. **Phase III:** Encode text, apply the mask in latent space, and decode to obtain an edited embedding for retrieval.

mean-pooled into a single vector. Queries and documents are ranked by cosine similarity between ℓ_2-normalized pooled embeddings (Karpukhin et al., 2020). In multilingual information retrieval (MLIR), the candidate pool mixes multiple languages, and language identity encoded in embeddings can inflate similarity for same-language pairs, leading to same-language crowding in the top ranks. LANGSAE EDITING mitigates this effect by first rewriting pooled embeddings into an overcomplete *sparse* feature representation, where each embedding is expressed by a small set of active latent units. This representation makes language-related signal identifiable as consistently activated units within each language, enabling selective suppression without applying a single global transformation uniformly to all inputs. The edited representation is then decoded back to the original embedding dimensionality, so retrieval remains a drop-in replacement for existing cosine similarity search and vector database infrastructure (Li et al., 2020; Huang et al., 2021) (Figure 1). Because the transformation operates purely on vectors, it can be applied both to retrofit stored document embeddings offline and to transform query embeddings at runtime, without modifying the underlying encoder or requiring access to raw text.

3.1 Phase I: Training LANGSAE on pooled embeddings

Let $\mathbf{e} \in \mathbb{R}^d$ denote the *raw* pooled embedding produced by a frozen base encoder for a text segment. LANGSAE is trained directly on raw pooled embeddings. At retrieval time, reconstructed and edited embeddings are ℓ_2-normalized before cosine similarity scoring (Section 4.1).

LANGSAE is an overcomplete sparse autoencoder with encoder E_θ and decoder D_ϕ, where the latent dimensionality is $m \gg d$; sparsity encourages reusable latent units and makes activation-frequency statistics meaningful for isolating language-associated features. Given $\mathbf{e}$, the encoder produces latent pre-activations, followed by a ReLU nonlinearity to obtain non-negative activations:

$$\mathbf{z} = \mathrm{ReLU}(E_\theta(\mathbf{e})) \in \mathbb{R}^m, \qquad z_i \geq 0 \ \ \forall i. \quad (1)$$

We impose sparsity by keeping only the top-k activated features per example (top-k applied directly to ReLU activations) (Makhzani and Frey, 2013). Let $\mathrm{TopK}(\mathbf{z}, k)$ return the indices of the k largest entries of $\mathbf{z}$. The sparsification operator $S(\cdot)$ is:

$$[S(\mathbf{z})]_i = \begin{cases} z_i, & i \in \mathrm{TopK}(\mathbf{z}, k), \\ 0, & \text{otherwise}. \end{cases} \quad (2)$$

We denote the resulting sparse code by

$$\tilde{\mathbf{z}} = S(\mathbf{z}) = S(\mathrm{ReLU}(E_\theta(\mathbf{e}))). \quad (3)$$

LANGSAE is trained to reconstruct pooled embeddings with mean squared error:

$$\mathcal{L}_{\text{rec}}(\theta, \phi) = \mathbb{E}_{\mathbf{e}} \left[\|\mathbf{e} - D_\phi(\tilde{\mathbf{z}})\|_2^2 \right]. \quad (4)$$

We additionally include sparsity-related auxiliary terms to encourage stable sparse features.

3.2 Phase II: Identifying language-associated latent features

After training, we identify language-associated latent units using activation statistics computed on a language-labeled probe set. Let $\mathcal{P}_\ell$ be probe texts in language ℓ. For each $t \in \mathcal{P}_\ell$, we compute its pooled embedding $\mathbf{e}(t)$ and sparse code $\tilde{\mathbf{z}}(t)$ (Eq. 3). We treat a latent unit as *active* if its sparse activation is non-zero (after top-k sparsification), and estimate how often each unit activates in each language:

$$p_{i,\ell} \;=\; \mathbb{E}_{t \sim \mathcal{P}_\ell}[\mathbb{I}(\tilde{z}_i(t) > 0)]\,. \tag{5}$$

Given a global threshold $\tau \in (0, 1]$, we construct three sets of units:

- **Frequent for language ℓ:** $\mathcal{F}_\ell(\tau) = \{i : p_{i,\ell} \geq \tau\}$.
- **Language-unique for ℓ:** $\mathcal{U}_\ell(\tau) = \{i \in \mathcal{F}_\ell(\tau) : \max_{\ell' \neq \ell} p_{i,\ell'} < \tau\}$.
- **Overlapping:** $\mathcal{O}(\tau) = \{i : i \in \mathcal{F}_\ell(\tau) \cap \mathcal{F}_{\ell'}(\tau)$ for some $\ell \neq \ell'\}$.

Intuitively, $\mathcal{U}_\ell(\tau)$ contains units that fire reliably for language ℓ but not for other languages, while $\mathcal{O}(\tau)$ contains units that are frequent across multiple languages, which may reflect shared scripts, tokenization regularities, or multilingual corpus artifacts. We keep these sets distinct to support different masking strategies at inference (Section 3.3).

The threshold τ controls how conservatively units are selected. Values near $\tau \approx 1.0$ retain only the most consistently activated units, while lowering τ can rapidly increase the frequent sets and risk including broadly used units. We compute $p_{i,\ell}$ on the held-out validation split used during LANGSAE training (Appendix A.3) and report a sensitivity sweep over τ in Appendix D. We also evaluate whether including overlapping units in the suppression set is beneficial (Appendix E).

3.3 Phase III: LANGSAE EDITING at inference

LANGSAE EDITING requires a language label ℓ for each embedding. In our benchmarks, ℓ is provided by the dataset. In deployed systems, ℓ can come from document metadata or a standard language identification module.

Given a pooled embedding $\mathbf{e}$ and its language ℓ, we first compute its sparse latent code $\tilde{\mathbf{z}}$ using the trained LANGSAE encoder and the top-k sparsification defined in Section 3.1. We then remove language-associated signal by masking a selected set of latent units derived from the activation-frequency statistics in Section 3.2. We consider two masking strategies: (i) *Unique-only*, masking $\mathcal{U}_\ell(\tau)$, and (ii) *Unique+Overlapping*, masking $\mathcal{U}_\ell(\tau) \cup \mathcal{O}(\tau)$. We use *Unique+Overlapping* in our main experiments (with $\tau = 0.999$) and report *Unique-only* as an ablation (Appendix E).

Concretely, LANGSAE EDITING applies the following steps:

1. **Encode to sparse features:** compute $\tilde{\mathbf{z}}$ from $\mathbf{e}$ using the trained encoder and top-k sparsification.
2. **Language-conditioned masking:** form a mask set $\mathcal{S}_\ell$ (either $\mathcal{U}_\ell(\tau)$ or $\mathcal{U}_\ell(\tau) \cup \mathcal{O}(\tau)$), then set the corresponding latent coordinates to zero to obtain a masked code $\tilde{\mathbf{z}}'$.
3. **Decode back to the original space:** reconstruct an edited embedding $\tilde{\mathbf{e}} = D_\phi(\tilde{\mathbf{z}}')$.
4. **Normalize for cosine scoring:** output $\bar{\tilde{\mathbf{e}}} = \tilde{\mathbf{e}}/\|\tilde{\mathbf{e}}\|_2$.

Masking can reduce the number of active features below k for some examples. We keep this behavior (rather than refilling from lower-ranked activations) to avoid reintroducing correlated units.

For retrieval, we apply the same transformation (Baseline, SAE Reconstructed, or LANGSAE EDITING) to both queries and documents, and rank candidates by cosine similarity between the resulting ℓ_2-normalized vectors:

$$s(q, d) = \langle \bar{\mathbf{e}}_q, \bar{\mathbf{e}}_d \rangle\,. \tag{6}$$

3.4 Control: SAE reconstruction without masking

To isolate the contribution of feature suppression from reconstruction, we define an SAE reconstruction control that passes embeddings through LANGSAE without masking:

$$\bar{\tilde{\mathbf{e}}} = \frac{D_\phi(\tilde{\mathbf{z}})}{\|D_\phi(\tilde{\mathbf{z}})\|_2}, \qquad \tilde{\mathbf{z}} = S(\mathrm{ReLU}(E_\theta(\mathbf{e}))). \tag{7}$$

Comparing **Baseline**, **SAE Reconstructed**, and **LANGSAE EDITING** in Section 4 separates the effect of sparse autoencoding from the effect of targeted language-feature suppression.

4 Experiments

4.1 Experimental Setup

Task. We evaluate mixed-language multilingual information retrieval (MLIR): each query retrieves from a single multilingual pool that contains documents from multiple languages, and relevant evidence may appear in any language.

Benchmarks. We use Belebele (Bandarkar et al., 2024) and XQuAD (Artetxe et al., 2020) on 10 languages: Arabic, Chinese, English, French, Hindi, Italian, Japanese, Portuguese, Russian, and Spanish. Both benchmarks provide parallel/aligned documents across languages (Belebele via passage IDs, XQuAD via SQuAD example IDs); each passage/context paragraph is treated as one retrieval unit. More details about benchmarks can be found in Appendix B.

Pools and relevance. For each benchmark we form a multilingual pool by taking the union of documents across the included languages, and we evaluate queries grouped by query language against the same shared pool. Belebele yields 4,880 documents (488 × 10) and 9,000 queries (900 × 10). XQuAD yields 1,440 documents (240 × 6) and 7,140 queries (1,190 × 6) for the available 6 languages (Arabic, Chinese, English, Hindi, Russian, Spanish). For each query q, the multi-relevant set R_q is the aligned document set across languages, so $|R_q| = 10$ on Belebele and $|R_q| = 6$ on XQuAD.

Retrieval and systems. Documents are not chunked at evaluation time. We encode text with mean pooling (Section 3), ℓ_2-normalize embeddings, and rank by cosine similarity using exact (brute-force) search over the full pool. We compare: (i) **Baseline**, frozen encoder; (ii) **SAE Reconstructed**, LANGSAE without masking; and (iii) **LANGSAE EDITING**, masking with the *Unique + Overlapping* strategy at τ=0.999. The same transformation is applied to both queries and documents.

Selecting τ. We compute activation frequencies $p_{i,\ell}$ on the held-out validation split used in LANGSAE training data preparation (Appendix A.2, A.3) and choose a conservative τ to avoid over-masking; we use $\tau = 0.999$ unless stated otherwise (Appendix D).

Metrics. We report standard Recall@20 and nDCG@20 (Järvelin and Käkäläinen, 2002) with binary relevance. In our mixed-language setting, each query q has a multi-relevant set R_q consisting of all aligned passages across languages (10 for Belebele, 6 for XQuAD). A retrieved passage is counted as relevant if it belongs to R_q. Recall@20 is the fraction of R_q retrieved in the top 20. nDCG@20 is computed with binary gains and an ideal ranking that places all relevant passages first (up to 20). We report averages per query language and a macro-average that weights each query language equally.

Retrieved Lang.	Avg. Count @ Top-20: m-e5-large	Avg. Count @ Top-20: LangSAE	Δ Count
Query Language (Bias Source)			
Chinese	**16.962**	**5.320**	**- 11.642**
Other Languages (Multilingual Targets)			
Arabic	0.000	0.321	+ 0.321
English	0.001	1.313	+ 1.312
Spanish	0.003	0.736	+ 0.732
Hindi	0.001	0.802	+ 0.801
Russian	0.102	1.122	+ 1.020
French	0.001	0.662	+ 0.661
Italian	0.067	0.972	+ 0.906
Japanese	0.003	0.811	+ 0.808
Portuguese	0.216	1.112	+ 0.897
Non-Zh Total	**0.394**	**7.852**	**+ 7.458**

Table 1: Ground-truth removal reveals same-language preference. Avg. retrieved language counts for Chinese queries (k=20, 900 queries).

4.2 Bias Evidence: Quantifying Language Bias via Ground-Truth Removal

To isolate same-language preference from semantic relevance, we measure the *language distribution of retrieved distractors* (Yang et al., 2021, 2025). Standard retrieval metrics can obscure language bias in our setting because multiple aligned ground-truth documents exist across languages, and a system may retrieve some ground-truth items while still allocating many remaining ranks to same-language non-relevant passages. To focus on this issue, we remove aligned ground-truth documents from the *retrieved list* before computing language counts.

Concretely, for each query we first retrieve the top-20 documents from the full multilingual pool. We then remove all aligned ground-truth documents for that query (across all included languages for that benchmark) from the retrieved list, and compute the number of remaining retrieved documents per language. After this removal, the retained documents are non-relevant under the benchmark labels, so their language distribution reflects lan-

Language	multilingual-e5-large		All-but-the-Top		SAE Reconstructed		LangSAE Editing	
	nDCG@20	Recall@20	nDCG@20	Recall@20	nDCG@20	Recall@20	nDCG@20	Recall@20
Belebele								
Arabic	0.4853	0.4750	0.4194	0.3909	0.4930	0.4844	**0.6810**	**0.6719**
English	0.7322	0.7246	0.7087	0.7028	0.7370	0.7298	**0.7635**	**0.7461**
Spanish	0.6857	0.6600	0.6692	0.6458	0.6888	0.6633	**0.7500**	**0.7288**
Hindi	0.3836	0.3329	0.3727	0.3209	0.3884	0.3393	**0.4483**	**0.4103**
Russian	0.2738	0.1794	0.2631	0.1674	0.2749	0.1804	**0.2766**	**0.1960**
Chinese	0.3397	0.2649	0.4175	0.3728	0.3461	0.2731	**0.6947**	**0.6821**
French	0.6847	0.6580	0.6842	0.6569	0.6918	0.6656	**0.7304**	**0.7099**
Italian	0.6522	0.6227	0.6416	0.6134	0.6603	0.6308	**0.7485**	**0.7276**
Japanese	0.5116	0.4769	0.5503	0.5344	0.5171	0.4836	**0.7119**	**0.7008**
Portuguese	0.6107	0.5634	0.6642	0.6229	0.6165	0.5712	**0.7292**	**0.7068**
Macro Average	0.5359	0.4958	0.5391	0.5028	0.5414	0.5022	**0.6534**	**0.6280**
XQuAD								
Arabic	0.6752	0.7557	0.6361	0.6975	0.6809	0.7632	**0.8362**	**0.8972**
English	0.8504	0.9147	0.8210	0.8829	0.8555	0.9199	**0.8751**	**0.9216**
Spanish	0.7838	0.8664	0.7991	0.8709	0.7884	0.8731	**0.8672**	**0.9198**
Hindi	0.7015	0.7751	0.7142	0.7777	0.7093	0.7840	**0.8443**	**0.8999**
Russian	0.7973	0.8908	0.7414	0.8284	0.8015	0.8936	**0.8956**	**0.9457**
Chinese	0.4765	0.4831	0.5854	0.6392	0.4819	0.4905	**0.8496**	**0.9080**
Macro Average	0.7141	0.7810	0.7162	0.7828	0.7196	0.7874	**0.8613**	**0.9154**

Table 2: MLIR performance on Belebele and XQuAD, reported by query language. We compare the base encoder, a global All-but-the-Top post-processing baseline, SAE reconstruction, and LangSAE Editing. Dark shading indicates the best result, light shading indicates the second best, computed per row and metric.

guage preference among distractors rather than the need to surface labeled answers. Because some of the original top-20 entries can be ground-truth documents that are removed from this analysis, the per-language counts in Table 1 are not expected to sum to 20. The difference to 20 equals the average number of ground-truth documents retrieved in the top-20 that were excluded from the distractor-only accounting.

We focus on Chinese queries on Belebele (900 queries). Table 1 shows clear evidence of same-language crowding among distractors. Under the baseline, Chinese accounts for 16.962 distractors on average, while all non-Chinese languages together account for only 0.394 distractors (17.356 distractors total). After applying LangSAE, the average number of Chinese distractors drops sharply to 5.320, while non-Chinese distractors increase to 7.852 (13.172 distractors total). In proportional terms, Chinese distractors drop from 97.7% of distractors (16.962/17.356) to 40.4% (5.320/13.172), while non-Chinese distractors rise from 2.3% to 59.6%. Since Belebele is parallel and the candidate pool is balanced across languages by construction, this shift is not explained by pool-size imbalance.

The gap to 20 also increases substantially, from $20 - 17.356 = 2.644$ in the baseline to $20 - 13.172 = 6.828$ under LangSAE. This indicates that LangSAE retrieves more aligned ground-truth items within the top-20 while simultaneously reducing same-language crowding among the remaining non-relevant candidates. Together, these results provide direct diagnostic evidence that LangSAE Editing mitigates same-language preference in mixed-language retrieval pools. Qualitative retrieval examples are provided in Appendix F.

4.3 MLIR Retrieval Performance

Table 2 summarizes retrieval quality by query language on Belebele and XQuAD under our mixed-language MLIR setting, where every query retrieves from the same multilingual pool. In addition to the base encoder (`multilingual-e5-large`), we include two post-hoc baselines to separate generic embedding-space post-processing from targeted language-identity suppression: All-but-the-Top (Mu et al., 2018), a global anisotropy-reduction transform that removes dominant principal components from the embedding space, and SAE Reconstructed, which passes embeddings through LangSAE without masking to isolate the effect of sparse autoencoding from feature suppression. We also observe consistent improvements when applying LANGSAE EDITING to a different multilingual embedding model (`jinaai/jina-embeddings-v3`), with results reported in Appendix C.

Overall, LANGSAE EDITING substantially outperforms both global post-processing and reconstruction-only controls. All-but-the-Top yields small or mixed changes across languages, which is expected because it applies a single language-agnostic linear projection and does not directly target language identity. Similarly, SAE reconstruction alone provides only marginal gains, indicating that improvements are not driven by generic reconstruction effects. In contrast, masking language-associated latent units produces consistent and often large gains, supporting the claim that retrieval improvements are driven by targeted suppression of language-identity features.

First, the gains are broad rather than isolated. Improvements appear across most query languages, indicating that the method is not merely fixing a small set of pathological cases. This supports the central claim that language identity acts as a systematic shortcut in similarity search: when language-associated signal inflates same-language similarity, it affects the ordering of many competitive candidates, not only a few outliers.

Second, gains concentrate in languages that are most separable by surface form. Languages with scripts or tokenization regimes that differ sharply from the Latin-script group tend to benefit the most. This is consistent with the mechanism we target. If the encoder embeds script and orthographic cues as easily recoverable language features, then the embedding space will naturally partition by language, and nearest-neighbor retrieval will spend much of its top-k capacity within the query-language region. By suppressing the latent units that behave like language identifiers, LANGSAE EDITING reduces this partitioning pressure and makes ranking depend more on shared semantic structure. The t-SNE projections in Section 4.4 qualitatively support this interpretation.

Finally, the improvements align with our bias-focused diagnostic in Section 4.2. After removing aligned ground-truth documents from the retrieved lists, the remaining distractors become less dominated by the query language, indicating that editing reduces same-language crowding among non-relevant candidates.

Method	Total (100k)	ms / sample	samples / s
SAE Reconstructed	3.1358 s	0.0314	31,889.37
LANGSAE EDITING	4.4516 s	0.0445	22,463.88
multilingual-e5-large	82.0601 s	0.8206	1,218.62

Table 3: Runtime of post-hoc embedding transformation vs. base encoding, measured over 100,000 samples. Timings measure GPU forward-pass compute only and exclude tokenization, disk IO, and ANN search, which are identical across methods.

Runtime and deployment efficiency. Table 3 compares the cost of post-hoc vector editing against re-running the base encoder. On 100,000 samples, LANGSAE EDITING takes 0.0445 ms per sample, while the base encoder takes 0.8206 ms, making editing $\approx 18.4\times$ cheaper than base encoding for corpus-wide updates and only a small overhead when applied after query encoding. Masking adds 0.0131 ms per sample over SAE reconstruction without masking. These results support our deployment claim that language-identity mitigation can be applied to stored embeddings offline and to queries at runtime with negligible compute compared to encoder tuning or corpus-wide re-encoding.

4.4 Visualizing Language Identity Isolation and Removal in Embedding Space

Figure 2 visualizes pooled embeddings from three representations using t-SNE (1000 samples per language): the base encoder space (left), an *inverse-mask* reconstruction that retains only the language-associated units (middle), and the LANGSAE EDITED space after suppressing those units (right). Each panel is produced by an independent t-SNE fit. We therefore interpret the plots qualitatively in terms of separation, overlap, and local neighborhood composition, rather than absolute distances or global geometry. Note that as with any 2D projection, t-SNE can distort distances and is sensitive to hyperparameters and random seeds, but it is useful for revealing dominant clustering structure.

Base embeddings exhibit strong language partitioning. In the base encoder space (left), points form visibly language-separated regions with limited overlap. This qualitative partitioning suggests that language identity is a salient organizing factor in the pooled embedding space. In a shared multilingual retrieval pool, such separation provides an intuitive mechanism for same-language crowding: if neighborhoods are predominantly monolingual,

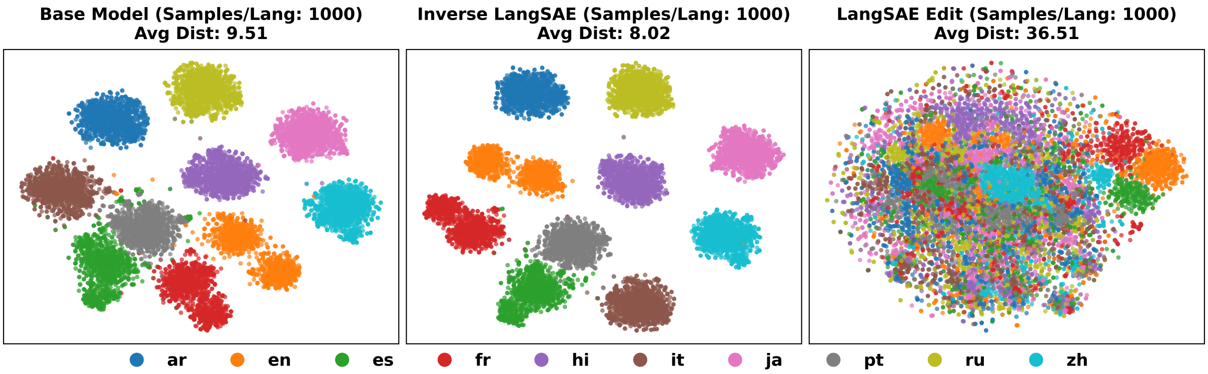

Figure 2: t-SNE projections of pooled embeddings (multilingual-e5-large, 1000 samples per language). **Left:** Base encoder embeddings. **Middle: Inverse mask** embeddings reconstructed using *only* the language-associated units. **Right:** LANGSAE EDITED embeddings after suppressing the language-associated units.

nearest-neighbor search can preferentially traverse within-language regions, making cross-language evidence less competitive even when it is semantically relevant.

Inverse masking isolates a language-identifying component. The inverse-mask visualization (middle) reconstructs embeddings using only the latent units identified as language-associated, while zeroing all other units. In this view, language separation remains pronounced and often appears sharper than in the base space. Qualitatively, this indicates that the selected latent units capture a concentrated component that is strongly predictive of language identity, and that this component alone is sufficient to recover clear language grouping in the embedding geometry.

LANGSAE EDITING reduces language-driven structure. After suppressing the language-associated units (right), the prominent language partitioning weakens and points from different languages interleave more substantially. While t-SNE does not preserve global distances, the visible increase in cross-language overlap in local neighborhoods is consistent with the intended effect of LANGSAE EDITING: reducing language identity as a shortcut signal so that similarity neighborhoods are less dominated by language membership and can be shaped more by semantic alignment.

Overall, Figure 2 provides a qualitative geometric view of the mechanism targeted by LANGSAE. The inverse-mask panel suggests that our activation-frequency selection isolates a language-identifying component, and the edited panel shows that suppressing this component reduces language-driven clustering, consistent with the reduced same-language crowding diagnostics (Section 4.2) and improved MLIR performance (Table 2).

5 Conclusion

We studied language bias in multilingual dense retrieval, where language identity encoded in embeddings can inflate similarity for same-language pairs and crowd out relevant evidence in other languages within a shared multilingual pool. We proposed LANGSAE, an overcomplete sparse autoencoder trained on pooled embeddings that identifies language-associated latent units via cross-language activation statistics. At inference time, LANGSAE EDITING suppresses these units and reconstructs edited embeddings in the original dimensionality, enabling retrofitting of existing vector databases without retraining the base encoder or re-encoding raw text. Experiments on mixed-language retrieval pools constructed from Belebele and XQuAD show consistent improvements in nDCG@20 and Recall@20 across languages, with diagnostic evidence that editing reduces same-language crowding among retrieved distractors. In addition to improving ranking quality, our results support a mechanistic view in which language identity occupies a small, controllable subset of sparse features that can be edited without broadly disrupting retrieval-relevant structure. Because the transformation is lightweight and vector-only, it can be applied both offline to update stored embeddings and at runtime as a small post-processing step on queries. These results indicate that language identity is a concentrated and editable factor in the representation space, and that targeted post-hoc suppression can improve MLIR in practical deployments where evidence may appear in any language.

Limitations

Our experiments primarily evaluate LANGSAE EDITING on two parallel multilingual QA benchmarks repurposed for mixed-language retrieval (Belebele and XQuAD). While this setting provides controlled alignment across languages and enables clean diagnostics of same-language crowding, it may not capture all properties of real-world multilingual corpora, such as domain shifts, uneven language distributions, or partially overlapping relevance across languages. The method assumes access to a language label for each embedding, which is available in our benchmarks but may require metadata or language identification in deployed systems. Finally, masking behavior is controlled by an activation-frequency threshold, and while we provide a sensitivity analysis, performance can degrade if suppression becomes too aggressive. Beyond language identity, embeddings can encode other correlated nuisance factors (e.g., script, domain, formatting), and suppressing language-associated features alone may not address all sources of bias or retrieval failures.

Ethics Statement

Our experiments use publicly available datasets and standard evaluation protocols. The underlying pretrained encoders may have been trained on large-scale web data that can contain biases, copyrighted material, or personal information beyond our control, and our work does not make claims about full pretraining data provenance. LANGSAE EDITING is intended to reduce same-language preference in multilingual retrieval, which can improve access to relevant information across languages, but it also changes the language and source distribution of retrieved results. In particular, increased cross-language retrieval may surface content that users cannot readily interpret or verify without translation, and downstream systems should consider whether to provide translation, provenance, or filtering to support safe use. Because our method is a post-hoc embedding transformation that can be applied at scale, practitioners should evaluate per-language behavior, monitor for unintended shifts in retrieval quality or exposure, and be transparent about the transformation when used in user-facing systems.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. volume 7, pages 597–610. MIT Press.

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.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the MS MARCO passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzm'an, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Xiyang Hu, Xinchi Chen, Peng Qi, Deguang Kong, Kunlun Liu, William Yang Wang, and Zhiheng Huang. 2023. Language agnostic multilingual information retrieval with contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9133–9146.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. Whiteningbert: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244.

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.

Zhiqi Huang, Hansi Zeng, Hamed Zamani, and James Allan. 2023. Soft prompt decoding for multilingual dense retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1208–1218. ACM.

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

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*. Also available as arXiv:1702.08734.

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. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 39–48. ACM.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130.

Jindřich Libovick'y, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual BERT? In *arXiv preprint arXiv:1911.03310*.

Jindřich Libovick'y, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674.

Alireza Makhzani and Brendan Frey. 2013. k-sparse autoencoders. *arXiv preprint arXiv:1312.5663*.

Yu. A. Malkov and D. A. Yashunin. 2020. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*.

Niklas Muennighoff, Nouamane Tazi, Lo"ic Magne, and Nils Reimers. 2022. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688.

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*.

Yuming Wu, Chaojun Xiao, Deyi Xiong, and Chong Yang. 2022. Unsupervised context aware sentence representation pretraining for multi-lingual dense retrieval. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI 2022)*, pages 4178–4185.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. 2022. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5617–5633.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jimmy Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations (ICLR)*.

Eugene Yang, Dawn J. Lawrie, and James Mayfield. 2024. Distillation for multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2368–2373. ACM.

Jinrui Yang, Fan Jiang, and Timothy Baldwin. 2025. Language bias in information retrieval: The nature of the beast and mitigation methods. *arXiv preprint arXiv:2509.06195*.

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021. A simple and effective method to eliminate the self language bias in multilingual representations. *arXiv preprint arXiv:2109.04727*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137. 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.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240.

A Training and Implementation Details

The following details specify the LANGSAE training configuration and the statistics used to compute activation-frequency features for language identification and masking.

A.1 Base encoder and pooled embeddings

We use `intfloat/multilingual-e5-large` as the frozen base encoder. For each input text segment, token representations are mean-pooled to obtain a single embedding $\mathbf{e} \in \mathbb{R}^d$ with $d = 1024$. LANGSAE is trained on raw pooled embeddings (no ℓ_2 normalization during training). At inference, reconstructed or edited embeddings are ℓ_2-normalized before cosine similarity scoring.

A.2 Training data construction

Sources and languages. The training corpus is constructed from mMARCO and MIRACL, restricted to 10 languages: Arabic, Chinese, English, French, Hindi, Italian, Japanese, Portuguese, Russian, and Spanish.

Tokenizer and segment length. All lengths are computed using the base encoder tokenizer. Examples with tokenized length below 250 tokens are discarded.

Length-based splitting. The goal is to expose LANGSAE to language-identity patterns in pooled embeddings, so a length-based segmentation scheme is used. For an example with tokenized length L:

LANGSAE checkpoint configuration	
Base encoder	`multilingual-e5-large`
Training corpora	mMARCO + MIRACL
Embedding dimension d	1024
Expansion factor m/d	256
Dictionary size m	262,144
Sparsity (top-k)	4,096
Learning rate	5×10^{-4}
Auxiliary loss coefficient	1×10^{-1}
Auxiliary usage target	2×10^{-2}
Epochs	1
Training and validation summary	
Aux loss (train / val)	0.9594 / 0.9719
Dead features (%) (train / val)	0 / 0
FVU (train / val)	0.002116 / 0.002109
ℓ_0 active features (train / val)	3454.35 / 3457.18
Total loss (train / val)	5.60×10^{-7} / 5.58×10^{-7}
MSE loss (train / val)	5.60×10^{-7} / 5.58×10^{-7}

Table 4: Top: LANGSAE checkpoint configuration used in main experiments. Bottom: training and validation summary for the same run.

- If $250 \le L \le 500$, keep it as a single segment.
- If $500 < L \le 1000$, take the first 500 tokens and split into two non-overlapping 250-token segments. Any remaining suffix shorter than 250 tokens is discarded.
- If $L > 1000$, partition into consecutive non-overlapping 500-token segments. Any remaining suffix shorter than 250 tokens is discarded.

All retained segments therefore fall within 250–500 tokens, and long examples yield multiple segments.

Balancing across languages. After chunking, segment counts differ across languages due to corpus variation. Each language is downsampled to match the smallest per-language segment count within each split, yielding balanced training and validation sets.

Train and validation sizes. After filtering, chunking, and balancing, the dataset contains **95,744,230** training segments and **23,936,060** validation segments. Training uses **1 epoch** over the training set.

A.3 Probe set for activation-frequency statistics

To compute activation frequencies $p_{i,\ell}$ for language-feature identification (Section 3.2), we use the same validation split described above, grouped by language. Specifically, we embed each validation segment with the frozen encoder, compute its sparse code $\tilde{\mathbf{z}}$, and estimate $p_{i,\ell} = \mathbb{E}_{t\sim\mathcal{D}_\ell}[\mathbb{I}(\tilde{z}_i(t) > 0)]$ using validation segments t in language ℓ.

A.4 Auxiliary feature-usage loss

In addition to reconstruction loss (Eq. 4), we employ an auxiliary feature-usage encouragement objective to mitigate the dead-feature problem in top-k sparse autoencoders. Intuitively, this loss penalizes latent units whose estimated activation frequency falls below a user-defined minimum usage target, encouraging the model to utilize the full dictionary capacity instead of collapsing onto a small subset of features.

Concretely, we define a per-unit *activation deficit* as the non-negative difference between a target activation fraction and the unit's estimated activation frequency. Activation frequencies are estimated differentiably (via a high-temperature sigmoid surrogate), and the auxiliary loss is computed as the mean squared activation deficit across units. The auxiliary coefficient and target used in the main checkpoint are reported in Table 4.

A.5 Optimization, precision, and hardware

We optimize LANGSAE with Adam. Training uses mixed precision (fp16) on CUDA via `torch.cuda.amp.autocast` and `GradScaler` for numerical stability. Training was performed on 8 NVIDIA RTX A6000 GPUs. For the expansion-factor-256 configuration used in our main experiments, training took approximately 6 hours.

A.6 LANGSAE training summary

Table 4 reports the configuration and logged statistics for the LANGSAE checkpoint used in the main experiments. The overcomplete dictionary size is determined by the expansion factor (m/d), and sparsity is enforced by top-k selection on ReLU activations (Section 3.1). Reported training statistics include the auxiliary loss used to encourage feature usage, the dead-feature rate, and reconstruction quality measured by fraction of variance unexplained (FVU). The ℓ_0 statistic corresponds to the number of non-zero latent activations per example after top-k sparsification.

B Evaluation Benchmarks

We evaluate multilingual retrieval in mixed-language pools using multilingual question answering (QA) datasets with parallel constructions, repurposed as retrieval tasks. These datasets provide aligned query and document instances across languages, enabling controlled cross-language evaluation without relying on heuristic relevance transfer. Since the original task is extractive QA, the associated passages serve as precise gold evidence for retrieval: for each question, the passage paired with that question is treated as relevant, and parallel variants of that passage across languages define aligned relevant evidence under our multilingual pool setting. This evaluation paradigm is widely used in recent work to assess multilingual and cross-lingual retrieval behavior using QA resources with parallel structure.

Belebele. Belebele (Bandarkar et al., 2024) is a professionally translated multilingual QA dataset designed to support high-quality multilingual evaluation across a broad set of languages. Translations were produced by native speakers proficient in English, aiming to preserve both contextual meaning and language-specific nuances. In our retrieval formulation, we treat each passage as a document and each question as a query. Because passages are parallel across languages via passage identifiers, we can construct a multilingual candidate pool by taking the union of passages across languages and define a multi-relevant set for each query consisting of all aligned passages across the included languages. This parallel structure makes Belebele well-suited for diagnosing same-language preference and measuring whether a retriever surfaces semantically aligned evidence across languages in a shared multilingual pool.

XQuAD. XQuAD (Artetxe et al., 2020) is a multilingual QA benchmark derived from SQuAD 1.1 (Rajpurkar et al., 2016). It provides translations of question-answer pairs and context paragraphs into multiple languages, yielding fully parallel examples across languages. In our retrieval formulation, each translated context paragraph is treated as a document and each translated question is treated as a query. The strict one-to-one alignment across languages enables constructing multilingual pools and defining aligned relevant document sets analogously to Belebele. This makes XQuAD a useful benchmark for evaluating the sta-

Language	jina-embeddings-v3		SAE Reconstructed		LangSAE Editing	
	nDCG@20	Recall@20	nDCG@20	Recall@20	nDCG@20	Recall@20
			Belebele			
Arabic	**0.5504**	0.5574	0.5450	0.5530	0.5474	**0.5586**
English	0.5806	**0.5876**	0.5836	0.5874	**0.5839**	0.5873
Spanish	0.6649	0.6739	0.6624	0.6690	**0.6653**	**0.6740**
Hindi	0.3752	0.3656	0.3782	0.3666	**0.3802**	**0.3683**
Russian	0.2071	0.1760	0.2103	0.1769	**0.2111**	**0.1774**
Chinese	0.5529	0.5430	0.5611	0.5538	**0.5648**	**0.5617**
French	0.6112	**0.6186**	0.6108	0.6159	**0.6134**	**0.6186**
Italian	0.6488	0.6529	0.6471	0.6510	**0.6520**	**0.6568**
Japanese	0.5786	0.5757	0.5767	0.5768	**0.5809**	**0.5828**
Portuguese	0.6318	**0.6388**	0.6304	0.6354	**0.6333**	**0.6388**
Macro Average	0.5401	0.5390	0.5405	0.5386	**0.5432**	**0.5423**
			XQuAD			
Arabic	0.7156	0.7894	0.7272	0.7976	**0.7302**	**0.8048**
English	0.7461	0.8169	**0.7558**	**0.8234**	0.7505	0.8221
Spanish	0.7927	0.8578	0.7980	0.8595	**0.8016**	**0.8651**
Hindi	0.7334	0.8003	0.7480	0.8113	**0.7509**	**0.8181**
Russian	0.7616	0.8293	0.7725	0.8373	**0.7742**	**0.8413**
Chinese	0.6950	0.7660	0.7171	0.7849	**0.7253**	**0.7952**
Macro Average	0.7408	0.8101	0.7530	0.8191	**0.7556**	**0.8247**

Table 5: MLIR performance on Belebele and XQuAD using jina-embeddings-v3, reported by query language. Dark shading indicates the best result, light shading indicates the second best, computed per row and metric.

bility of embedding-based retrieval under linguistic variation and for measuring how language-identity signal affects similarity search in multilingual settings.

Why parallel QA datasets. We require a benchmark construction where, for every query, there is guaranteed relevant evidence in multiple languages within a single shared candidate pool. This is essential for diagnosing *same-language crowding* cleanly: we need cross-language relevant documents to exist by construction so that a retrieval failure under a mixed-language pool can be attributed to language-identity effects rather than missing cross-language relevance labels or incomplete cross-language annotation. Parallel QA datasets provide this property through their one-to-one alignment across languages, allowing us to define multi-relevance sets that include all aligned passages for each query and to evaluate whether retrieval surfaces semantically matching evidence beyond the query language.

We considered a broader range of multilingual retrieval datasets, but many multilingual IR benchmarks are designed primarily for *monolingual* retrieval within each language and therefore do not provide fully parallel, one-to-one aligned query–document pairs across languages. In particular, Mr. TyDi (Zhang et al., 2021) and MIRACL (Zhang et al., 2023) contain language-specific query sets and relevance judgments over language-specific corpora, rather than a shared pool with guaranteed cross-language aligned relevant documents for every query. This makes it non-trivial to construct controlled multi-relevance sets in mixed-language pools without introducing additional cross-language alignment machinery (e.g., entity or document linking). Because our goal is to isolate and measure same-language preference in a setting where cross-language relevant evidence is present by construction, we focus on Belebele and XQuAD as our primary evaluation datasets.

C Additional Encoder Results: `jinaai/jina-embeddings-v3`

To evaluate whether LangSAE Editing generalizes beyond `multilingual-e5-large`, we repeat the mixed-language MLIR protocol from Section 4 using `jinaai/jina-embeddings-v3` as the frozen base encoder. We train a separate LangSAE on pooled embeddings produced by this encoder, then apply the same inference-time editing procedure to both query and document vectors. Unless stated otherwise, we use an expansion factor of 128 with top-k=2048, learning rate 3×10^{-4}, auxiliary coefficient 10^{-1}, and usage target 2×10^{-2}. At inference we use the Unique+Overlapping masking strategy with τ=0.999.

Table 5 reports nDCG@20 and Recall@20 by query language on Belebele and XQuAD. Overall, the post-hoc transformation yields consistent, if smaller, improvements compared to the base encoder, indicating that the language-identity signal exploited by similarity search is not specific to a single encoder family and that sparse feature suppression can provide benefits across encoder architectures.

D Sensitivity to Activation-Frequency Threshold τ

multilingual-e5-large				
	Belebele (Macro Avg)		XQuAD (Macro Avg)	
Threshold	nDCG@20	Recall@20	nDCG@20	Recall@20
1.000	0.5974	0.5717	0.7876	0.8669
0.999	0.6534	0.6280	0.8613	0.9154
0.998	0.6019	0.5767	0.8258	0.8775
0.997	0.4817	0.4615	0.7369	0.7843
0.996	0.2768	0.2708	0.5130	0.5640
0.995	0.0874	0.0939	0.2043	0.2505
0.990	0.0455	0.0481	0.1495	0.1742

Table 6: Sensitivity to activation-frequency threshold τ (absolute macro-average).

The activation-frequency threshold τ controls how conservatively LangSAE Editing selects latent units for suppression. Using per-language activation frequencies $p_{i,\ell}$ (Eq. 5), we define frequent sets as $\mathcal{F}_\ell(\tau) = \{i \mid p_{i,\ell} \geq \tau\}$, derive language-unique and overlapping sets $\mathcal{U}_\ell(\tau)$ and $\mathcal{O}(\tau)$ as in Section 3.2, and apply the same masking strategy used elsewhere:

$$\mathcal{S}_\ell(\tau) = \mathcal{U}_\ell(\tau) \cup \mathcal{O}(\tau). \tag{8}$$

Table 6 reports *absolute* macro-average nDCG@20 and Recall@20 under mixed-language retrieval in multilingual pools. Performance exhibits a narrow high-performing band near $\tau \approx$ 0.999–1.000: τ = 0.999 achieves the strongest results on both benchmarks, and $\tau \in \{1.000, 0.998\}$ remains competitive. However, once τ is relaxed further, performance degrades rapidly. By $\tau = 0.997$ metrics drop substantially, and by $\tau \leq 0.995$ retrieval quality collapses to very low values under our multi-relevance evaluation, indicating that masking has removed substantial retrieval-relevant structure.

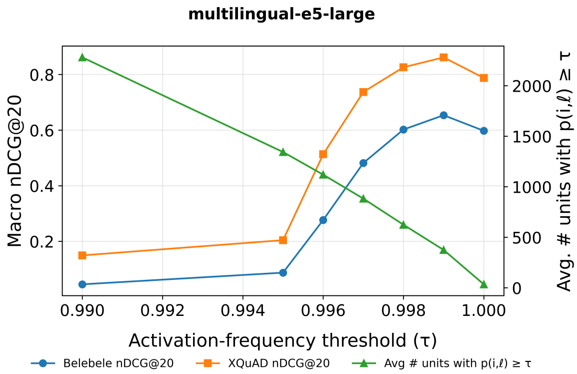

Figure 3: Sensitivity to τ. Macro nDCG@20 as a function of the activation-frequency threshold τ (left axis), together with the average number of latent units per language whose activation frequency exceeds τ (right axis). As τ decreases slightly below 1.0, the set of frequently active units grows rapidly, which propagates to a much larger suppression set and can trigger over-masking.

Figure 3 clarifies why small changes in τ can have outsized effects. The right axis shows that the number of latent units with $p_{i,\ell} \geq \tau$ increases sharply as τ decreases in the narrow region just below 1.0. This reflects a concentration of activation frequencies near one, which is expected in a top-k sparse autoencoder where a subset of features is reused consistently across many inputs. Because $\mathcal{S}_\ell(\tau)$ is built from frequent sets (and includes both language-unique and overlapping frequent units), this rapid growth propagates into a much larger suppression set. Past a critical point, masking begins to remove not only language-associated shortcut features but also frequently used factors that support semantic similarity, which contracts similarities for both positives and competitive negatives and destroys the relative separability required for accurate ranking.

Overall, these results show that activation-frequency thresholding must be used conservatively. Values in a tight neighborhood near $\tau \approx$ 0.999–1.000 can suppress highly consistent language-associated units while preserving most shared semantic structure, whereas modest additional relaxation triggers over-masking and severe degradation.

E Overlap Removal vs. Non-Removal

Overlap handling controls whether LANGSAE EDITING suppresses only language-unique features, or suppresses both language-unique and overlapping features that are frequent across multiple languages. The comparison uses three settings: (i) no feature suppression (Baseline), (ii) suppression of language-unique features only, and (iii) suppression of language-unique plus overlapping features. Table 7 reports macro-average results on Belebele under the same evaluation protocol as the main experiments.

Including overlapping features in the suppression set yields the best overall performance, while masking language-unique features alone can degrade retrieval relative to the baseline. This suggests that language bias in the embedding space is not carried exclusively by language-unique units. Instead, a substantial portion of the shortcut signal appears to live in features that are frequent across multiple languages, for example shared script- or tokenization-related regularities, or multilingual corpus artifacts, that can still inflate cosine similarity and contribute to same-language crowding. Masking only language-unique units may therefore remove some retrieval-relevant variation without sufficiently attenuating the dominant cross-language shortcut features, whereas additionally suppressing overlapping features more effectively weakens language-driven similarity and improves ranking in mixed-language pools.

F Qualitative Ranking Examples

To complement the aggregate metrics, we present two qualitative examples that show how LANGSAE EDITING changes the composition of the top-ranked results in mixed-language pools. In both cases, the baseline retrieval list is dominated by same-language items, and several of these high-ranked candidates are distractors that are only weakly related to the query. After editing, the ranking surfaces additional aligned evidence written in other languages, increasing cross-language coverage while preserving the ability to retrieve relevant passages in the query language. These examples align with our quantitative analysis of same-language crowding (Table 1).

multilingual-e5-large		
	Average	
Removal Strategy	**nDCG@20**	**Recall@20**
No Removal	0.5359	0.4958
Unique Only	0.5176	0.4696
Unique + Overlapping	**0.6534**	**0.6280**

Table 7: Macro-average results on Belebele under different suppression strategies. Suppressing overlapping features in addition to language-unique features yields the strongest MLIR performance.

Markers indicate whether the retrieved passage is aligned ground-truth evidence for the query (**O**) or not (**X**).

Query (ZH): 根据这段文字,亚马逊河的河水来自哪里?

English: Based on this text, where does the water of the Amazon River come from?

Rank	multilingual-e5-large	Rel.	LANGSAE	Rel.
1	亚马逊河是世界上第二长,也是最大的河流 它的水量是第二大河流的 8 倍以上...	O	亚马逊河是世界上第二长,也是最大的河流 它的水量是第二大河流的 8 倍以上...	O
2	1963 年大坝建成后,季节性洪水被控制住了,沉积物不再冲散到河流里...	X	[PT] O Amazonas é o maior rio e o segundo mais longo da Terra...	O
3	维京人利用俄罗斯水路到达黑海和里海 其中一些路线至今仍可通行...	X	[EN] The Amazon River is the second longest and the biggest river...	O
4	印度河流域文明是青铜时代的文明,位于印度西北部次大陆...	X	[FR] Le fleuve Amazone est le deuxième plus long et le plus grand...	O
5	联合国维和人员在 2010 年地震后抵达海地,他们因疫情蔓延而受到指责...	X	[ES] El río Amazonas es el más caudaloso y el segundo más extenso...	O

Table 8: Qualitative retrieval example (Amazon River query). **O** indicates the passage contains the correct evidence, **X** otherwise.

Query (ES): Según el texto, ¿cuál de las siguientes opciones no se recomienda para que los atletas jóvenes disfruten más el deporte?

English: According to the text, which of the following is not recommended for young athletes to enjoy sports more?

Rank	multilingual-e5-large	Rel.	LANGSAE	Rel.
1	No es posible que las prácticas nutricionales adecuadas, por sí solas, generen un rendimiento de elite...	O	No es posible que las prácticas nutricionales adecuadas, por sí solas, generen un rendimiento de elite...	O
2	[PT] A nutrição adequada por si só não gera desempenhos de alta performance, mas pode afetar...	O	[PT] A nutrição adequada por si só não gera desempenhos de alta performance, mas pode afetar...	O
3	La carrera de distancia media es un deporte relativamente económico; no obstante...	X	La carrera de distancia media es un deporte relativamente económico; no obstante...	X
4	USA Gymnastics respalda la nota del Comité Olímpico de los Estados Unidos...	X	[ZH] 仅靠适当的营养实践并不足以造就出色表现,但这可以显著影响年轻运动员...	O
5	El ganador olímpico de la medalla de oro debía nadar en el estilo libre de 100 metros...	X	[IT] Le sole pratiche nutrizionali corrette non bastano a generare elevate prestazioni...	O

Table 9: Qualitative retrieval example (athlete nutrition query). **O** indicates the passage contains the correct evidence, **X** otherwise.