ETBENCH: EVALUATING EMBEDDING TRANSLATION FOR MIXED-INDEX RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixed-index retrieval searches documents encoded by different embedding models, *e.g.,* a legacy model and its replacement. Embedding translation enables this by mapping stored source embeddings into the target space, where queries are also encoded. Translation often uses embedding alignment, as in bilingual lexicon induction, fitting a map to anchor pairs, *i.e.,* the same documents encoded by both models, and applying it to the remaining source-only vectors. We show that better alignment does not necessarily yield better mixed-index retrieval, and prove a sufficient and necessary condition under which alignment maps can preserve the target model's top-k set. We further develop ETBench, a benchmark evaluating alignment methods for embedding translation across 24 directed encoder pairs and four corpora. It tests whether translation makes source-only documents searchable alongside native target embeddings in a mixed index. Our results show that high alignment accuracy can coexist with poor retrieval of translated documents, and common retrieval measures can be insensitive to this.

1 INTRODUCTION

In vector search, an embedding model represents documents and queries as vectors, and a query returns relevant documents whose vectors are closest to the query vector. Replacing the model, or merging vector databases built with different models, requires re-encoding every document, which is costly at scale and impossible if the original documents are unavailable. *Embedding translation* instead maps stored source vectors into the target model's space (Yoon & Arik, 2025; Yang et al., 2025; Yang & Cao, 2026). A map is learned from anchors, which are documents encoded by both models, and applied to the source-only documents, transforming them into the target model's embedding space; the target's queries then search these translations alongside native vectors encoded by the target itself, using distances of the target's embedding space. We call this *mixed-index retrieval*. Each translation stands in for a target encoding that was never actually produced by the target model.

The map usually comes from *embedding alignment*, the classic technique that brings corresponding vectors of two spaces together (Mikolov et al., 2013; Lample et al., 2018; Ruder et al., 2019), as in bilingual lexicon induction, which maps word vectors across languages to recover word translations. Alignment is judged on held-out pairs: an alignment map is accurate if it carries an unseen source vector nearest to that document's own target encoding. Since embedding translation reuses these maps, one expects better alignment to give better translation and hence more accurate mixed-index retrieval. In this work, we question this expectation and ask exactly when the two agree or disagree.

Our central proposition is that *better alignment need not yield better translation for mixed-index retrieval*. The objectives differ. Alignment asks whether a mapped vector lands nearer to its own target encoding than to any other document's. Mixed-index retrieval asks whether, for a query, that vector keeps its place against native vectors that carry no translation error. An error too small to change the recovered counterpart for alignment can still change a candidate's ranking for a query.

We make this distinction precise with a tight condition under which alignment-error bounds guarantee preservation of the target model's top-k results. The condition compares errors with query-dependent ranking gaps, and is both sufficient and necessary. We also show that maps with arbitrarily small alignment errors can have drastically different retrieval performance over the translated mixed index.

To understand how real embedding translators work, we build ETBench, a benchmark for mixed-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

index retrieval. It compares the translated mixed index with the same documents encoded entirely by the target model, and measures retrieval of translated and native documents separately. Across 24 directed encoder pairs and four corpora, we find that the alignment method with highest accuracy is not the best translator, and common retrieval measures can conceal poor retrieval of translated documents. Further, the gap between alignment map and downstream tasks over its translated vectors goes beyond mixed-index retrieval, applying to ***e.g.,*** KV cache translation as well.

In summary, we make the following contributions.

- *Theory*. We give a necessary and sufficient condition under which alignment-error bounds preserve mixed-index retrieval, and show why high alignment scores alone do not suffice (Section 2).
- *Benchmark*. We develop ETBench, a benchmark for embedding translation (Section 3).
- *Findings*. We show that high alignment accuracy can coexist with poor retrieval of translated documents, which common aggregate retrieval measures can conceal (Section 4).

2 From embedding alignment to embedding translation

We define embedding translation and ask when alignment errors preserve retrieval.

2.1 Embedding Translation and its Alignment Step

Pairwise embedding translation. Let f_s and f_t be source and target encoders with output dimensions d_s and d_t. We distinguish three disjoint document sets: $\texttt{D0}$ contains anchors encoded by both models; $\texttt{D1}$ contains source-only documents to translate; and $\texttt{D2}$ contains target-only documents. A *native* vector is one produced by f_t itself. A *translator* is a function $T : \mathbb{R}^{d_s} \to \mathbb{R}^{d_t}$ that predicts $f_t(d)$ from $f_s(d)$ for each $d \in \texttt{D1}$, using only the stored vectors $f_s(\texttt{D0} \cup \texttt{D1})$, $f_t(\texttt{D0} \cup \texttt{D2})$ and the anchor pairings, without accessing raw documents or invoking either encoder.

Translation via alignment. Translation usually obtains T from an alignment map fitted to the anchor pairs $(f_s(a), f_t(a))$, $a \in \texttt{D0}$. Alignment aims to bring corresponding representations together (Mikolov et al., 2013; Lample et al., 2018); the fitted map then applies to unseen source vectors in $f_s(\texttt{D1})$. If alignment uses shared coordinates, translation also needs a decoder into target coordinates. We use T for the complete conversion, including any decoding and the normalization below.

Mixed index. We store translated documents beside native ones:

$$I_T = \{(d, T(f_s(d))) : d \in \texttt{D1}\} \cup \{(d, f_t(d)) : d \in \texttt{D2}\}.$$

Each query is encoded by the target model in its query format, which may differ from its document format. Throughout this section, the query vector $q \in \mathbb{R}^{d_t}$, native vectors $f_t(d)$ and translated outputs $T(f_s(d))$ are nonzero and normalized. We rank all candidates by decreasing $q^\top v$, equivalently by increasing Euclidean or cosine distance, since $\|q - v\|_2^2 = 2 - 2q^\top v$.

The ideal translation reproduces the rankings of the hypothetical *all-native index* on the same documents: $I_{\text{native}} = \{(d, f_t(d)) : d \in \texttt{D1} \cup \texttt{D2}\}$. The missing $f_t(d)$ for $d \in \texttt{D1}$ is well-defined but unavailable to the translator. Both indexes exclude the anchors $\texttt{D0}$ and use the same queries, candidates and relevance; only $\texttt{D1}$'s representations differ. Exact reconstruction suffices for identical rankings but is not necessary, and the all-native ranking need not be optimal for relevance.

2.2 Why better alignment need not give better translation

We compare two requirements on the same *held-out* documents $\texttt{D1}$, whose target encodings are withheld from fitting. Alignment asks whether their mapped vectors recover their own target counterparts; translation asks whether target queries retrieve them alongside native documents.

Alignment. For $d \in \texttt{D1} \cup \texttt{D2}$, write $z_d = f_t(d)$ for its native vector and v_d for its stored vector: $v_d = T(f_s(d))$ on $\texttt{D1}$ and $v_d = z_d$ on $\texttt{D2}$. Its *translation error* is $e_d = v_d - z_d$. On $\texttt{D1}$, the *counterpart cosine* is $z_d^\top v_d$. Counterpart recovery asks whether z_d is the nearest vector to v_d among $f_t(\texttt{D1})$; the fraction of correct matches is $\texttt{Hits@1}$. These comparisons use the hypothetical target encodings, not additional inputs to the translator or mixed index.

Since the vectors are unit, $z_d^\top v_d = 1 - \|e_d\|_2^2/2$: counterpart cosine determines error length, but not direction. Let ρ_d be the distance from z_d to the nearest other vector in $f_t(\texttt{D1})$. The condition $\|e_d\|_2 < \rho_d/2$ suffices for correct recovery, since every other vector is at least $\rho_d - \|e_d\|_2 > \|e_d\|_2$ away from v_d. Thus perfect `Hits@1` allows nonzero errors; it does not imply exact reconstruction.

Translation. For a target query q, write the native and mixed-index scores as $s(q,d) = q^\top z_d$ and $\widehat{s}(q,d) = q^\top v_d$. For two candidates $d, b \in \texttt{D1} \cup \texttt{D2}$, their score difference is

$$\widehat{s}(q,d) - \widehat{s}(q,b) = \underbrace{q^\top(z_d - z_b)}_{\text{native score margin}} + \underbrace{q^\top(e_d - e_b)}_{\text{change caused by translation}} . \tag{1}$$

Suppose a translated document $d \in \texttt{D1}$ ranks above a native document $b \in \texttt{D2}$ with native score gap $\gamma = s(q,d) - s(q,b) > 0$. Since $e_b = 0$, d remains ahead precisely when $q^\top e_d > -\gamma$. Alignment error, *i.e.,* counterpart cosine between native and translated vectors for d, determines $\|e_d\|_2$, whereas retrieval depends on the projection $q^\top e_d$ relative to the ranking gap. Even with a correct held-out counterpart match, an irrelevant native document can displace a relevant translated one from the top k.

2.3 When does alignment suffice for translation?

We now characterize when alignment-error bounds guarantee preservation of the all-native top-k set, and why high alignment accuracy alone cannot give a guarantee across workloads.

From alignment error to retrieval correctness. Fix a unit target query q and $N = |\texttt{D1} \cup \texttt{D2}|$ unit native document vectors in $\mathbb{R}^{d_t}$, with $d_t \geq 2$ and $1 \leq k < N$. Write $q(I)$ for the top-k answer set returned by index I, and let $R_k = q(I_{\text{native}})$ have a strict boundary, *i.e.,* no score tie between this set and the rest. Further, we also assume that the source vectors $f_s(d)$, $d \in \texttt{D1}$, are pairwise distinct. Write $\theta_d = \arccos(q^\top z_d)$; hence, a smaller angle means a higher score.

For each document, let $\alpha_d \in [0, \pi]$ be an angular-error bound, with $\alpha_d = 0$ on `D2`. We call T *α_d-bounded* if, for every $d \in \texttt{D1}$, $\arccos(z_d^\top v_d) = 2\arcsin(\|e_d\|_2/2) \leq \alpha_d$. These bounds concern alignment held-out errors over `D1`, not anchor-fitting errors over `D0`. Given any such bounds, we study arbitrary maps with unit outputs subject only to these bounds.

Theorem 1*(a) For any fixed query q, every α_d-bounded alignment map T has $q(I_{\text{native}})$ as its unique top-k set in I_T if and only if*

$$\alpha_d + \alpha_b < \theta_b - \theta_d \quad \textit{for every } d \in q(I_{\text{native}}),\ b \notin q(I_{\text{native}}). \tag{2}$$

(b) For every $\varepsilon > 0$ and $\rho \in (0, 1]$, there exist a query q, document collections `D0`, `D1` and `D2`, with `D0` of any prescribed size, and two orthogonal alignment maps T, T' that fit every anchor exactly and have perfect held-out `Hits@1`, whose angular alignment errors on `D1` equal β_T and $\beta_{T'}$ on every document, with $\beta_T = \rho\,\beta_{T'} \in (0, \varepsilon)$, yet $q(I_T) \cap q(I_{\text{native}}) = \varnothing$ and $q(I_{T'}) = q(I_{\text{native}})$.

Theorem 1 characterizes the connection between alignment and translation, and tells us the following.

(1) When alignment is sufficient. For $d \in R_k$ and $b \notin R_k$, translation can increase d's angle to the query by at most α_d and decrease b's by at most α_b. Their sum must be smaller than the native angular gap to keep d ahead in every allowed configuration. For a translated d and native b, this reduces to $\alpha_d < \theta_b - \theta_d$. If the condition fails, the bounds permit a tie or reversal, but a particular fitted map may still succeed. Preserving the top-k set preserves Recall@k; preserving its internal order as well suffices for identical nDCG@k. Appendix A proves the theorem and gives the ordered-list extension.

(2) High alignment scores $\not\Rightarrow$ good translation. Theorem 1 (b) constructs orthogonal maps that fit every anchor and recover every held-out counterpart, yet give disjoint mixed-index answers. With $\rho = 1$, the maps have identical counterpart cosines; with $\rho < 1$, the map whose answer is disjoint from the native answer has smaller angular error on every held-out document. Thus no positive error tolerance independent of the query and collection guarantees retrieval preservation. For any fixed query and collection with a strict boundary, however, Theorem 1 (a) guarantees preservation once the errors are sufficiently small. We must therefore compare held-out errors with query-dependent ranking gaps, not alignment scores alone. Appendix A gives the construction and its anchor-span limitation.

Multiple source encoders. The same condition applies to several source encoders mapped into one target space: with one stored vector per document and the assumptions above imposed on each map, we compare each pair's summed error bounds against its gap in the combined index.

Related work. Embedding translation has been motivated by retrieval during partial model updates, which is studied in FastFill (Jaeckle et al., 2023), metric-compatible backfilling (Seo et al., 2023), Embedding-Converter (Yoon & Arik, 2025), vector database integration (Yang et al., 2025; Yang & Cao, 2026), and transformation through classification (Hamideche et al., 2026). We complement these studies with an exact alignment-to-ranking characterization and a benchmark evaluation that separates alignment from the retrieval of translated and native embeddings mixed as a single index.

There has been a host of work on applications of embedding alignment like translation, including *e.g.,* dictionary induction in domain tasks (Glavaš et al., 2019), multilingual candidate pool by LAReQA (Roy et al., 2020), geometric agreement across models (Huh et al., 2024; Maystre et al., 2025), and gallery backfilling across model updates (Jaeckle et al., 2023; Seo et al., 2023). Specifically, it has been shown that an alignment loss can shrink the length of errors (Jaeckle et al., 2023) of downstream tasks like quantization (Guo et al., 2020), while leaving the direction free. These studies measure the alignment itself or all of its aligned output as a whole. Instead, we study its impact in a mixed index induced by embedding translation using a fitted alignment map, where a mapped vector competes with native vectors that never pass through alignment. Theorem 1 formalizes the question and answers it exactly, and Sections 3- 4 measure it via a systematic benchmark study.

Appendix B discusses benchmark efforts on alignment and retrieval. Here we focus on embedding translation and its induced mixed-index retrieval, which is different from alignment and plain retrieval.

3 ETBENCH: MEASURING TRANSLATION IN THE MIXED INDEX

Section 2 analytically separates alignment from translation. We next build ETBench, a benchmark for embedding translation to further consolidate this distinction. We start with the benchmark setup below.

Encoders and datasets. We evaluated embedding translation across 18 encoders over 4 BEIR document collections (Thakur et al., 2021). Each collection contains built-in queries and answers as groundtruth for evaluation. We cover 24 translation directions by varying model families and sizes. One encoder pair on one corpus is referred to as an *evaluation cell*, giving up to 96 cells per translation method. Details about the encoders and datasets are provided in Appendices C-D.

Evaluation protocol. We partition each corpus once into the three sets of Section 2.1, with nominal proportions of one half for anchors `D0` and one quarter each for `D1` and `D2`. Built-in answers to any query are kept outside the anchor pool. The same partition is used for every method and encoder pair. Translators may read the stored source vectors of $\texttt{D0} \cup \texttt{D1}$, target vectors of $\texttt{D0} \cup \texttt{D2}$ and anchor pairings, but not documents, encoders, queries or relevance labels. The evaluator alone holds $f_t(\texttt{D1})$, for counterpart tests and the all-native reference. Neither retrieval index contains anchors.

Following Section 2, queries are always encoded by the target encoder, and all embedding vectors are normalized and have the same unit length, so dot product, cosine and Euclidean distance give the same order. We search I_T and I_{native} with identical queries, candidates and benchmark answers. Only the representations of `D1` differ. Unless stated otherwise, the mixed index I_T contains one source part `D1` and pairwise translation from `D1` to `D2` is assumed by default.

Translators. We evaluate 9 translators and the `identity` control that inserts source vectors without a learned map. The translators differ in both their use of anchors and their alignment over `D0`:

- *Without anchor pairings.* `vec2vec` (Jha et al., 2025) learns a shared intermediate representation from unpaired vector sets, here disjoint halves of `D0`. Gromov–Wasserstein (`GW`) (Peyré et al., 2016; Alvarez-Melis & Jaakkola, 2018) matches within-space distances; our configuration uses 256 anchors and serves weighted averages of target anchors for new inputs.
- *Linear and anchor-based maps.* `procrustes` (Schönemann, 1966) fits an orthogonal map with a scale, and reconstructs target vectors. Canonical correlation analysis (`cca`) fits correlated linear coordinates and decodes them into the target space. Relative Encoding (`relative`) (Moschella et al., 2023) describes an input by its similarities to source anchors; our translator decodes this

Table 1: Overall results of 9 translators, over 24 ordered encoder pairs (Table A3) on 4 BEIR corpora (SciFact, NFCorpus, ArguAna, FiQA); every entry is the mean over these pairs and corpora. (a) Right of the vertical line are the translators' mean retention results per each metric (nDCG@10, MRR@10, Recall@100) for each group (full, D1, D2), so 1 is native quality. (b) Left are alignment fitting (D0) and held-out alignment (D1), where cos is the mean cosine between a translated vector and its native target encoding, and Hits@1 is the fraction of translated vectors whose nearest neighbour is their own target encoding. Best results are highlighted in **bold**; D0 and D2 columns carry no bold as D0 is the set every map was fitted on, and D2's ideal is 1.000, where a departure in either direction means translated and native vectors compete out of balance.

	Hits@1		cos		nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	D0	D1	D0	D1	full	D1	D2	full	D1	D2	full	D1	D2
identity	.253	.252	.176	.176	.664	.054	1.117	.713	.046	1.147	.651	.126	1.032
relative	.997	.520	.987	**.978**	.796	**1.001**	.661	.824	**.995**	.664	.866	1.072	.746
procrustes	.913	.898	.948	.934	.603	.293	.842	.649	.249	.890	.779	.868	.727
cca	.973	**.936**	.949	.918	.871	.908	.852	.883	.861	.874	.938	1.074	.863
GW	.849	.837	.803	.803	.593	.148	.927	.647	.123	.976	.657	.577	.737
mlp	.954	.906	.867	.831	.874	.977	.801	.892	.933	.820	.897	**1.090**	.782
vec2vec	.001	.001	.503	.503	.599	.004	1.045	.657	.004	1.082	.542	.037	.919
embconv	.950	.870	.879	.786	.883	.727	.994	.887	.688	1.012	.966	.965	.976
la2m	.893	.894	.831	.852	.941	.800	1.034	.935	.756	1.045	**.992**	.982	1.001
hmoe	.834	.812	.882	.859	**.945**	.965	.929	**.948**	.940	.933	.979	1.046	.942

description using the paired target anchors.

- *Predicting target vectors.* A multilayer perceptron (mlp) predicts target coordinates directly. Embedding-Converter (embconv) (Yoon & Arik, 2025) also penalizes errors in target distances between sampled document pairs and between neighbors. la2m (Yang et al., 2025) uses local maps fitted on clustered anchors; hmoe (Yang & Cao, 2026) selects specialized maps through a hierarchy.

Among them, we classify embconv, la2m, and hmoe as *translation-oriented* methods, as they account for retrieval with target-space vectors when fitting the alignment maps, the mlp serves as a generic supervised regression baseline, while the others are *alignment-oriented*, as their alignment fitting is agnostic to documents already encoded in the target space.

Measurements. We measure the quality of alignment and translation as follows.

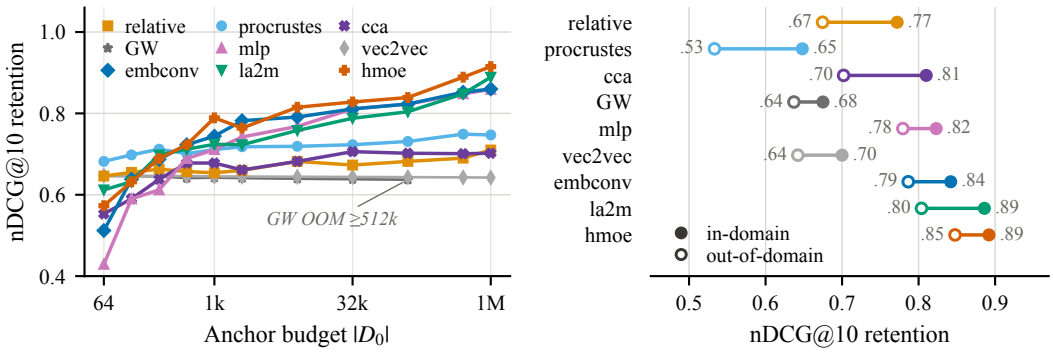
Held-out alignment. Alignment quality is measured over D1. For each translator, we evaluate its alignment map quality by calculating the mean counterpart cosine $z_d^\top v_d$ and Hits@1, which is the fraction of translations whose nearest vector among $f_t(\texttt{D1})$ is their own target encoding (Section 2.2).

Grouping. Translation quality is measured by retrieval quality over the induced mixed index I_T. However, instead of using a retrieval score aggregated over the entire dataset, we evaluate retrieval quality for three groups: (a) retrieve all built-in query-answer relevance judgments (*full*), (b) retrieve those in D1, and (c) retrieve those in D2. For group D1, we keep queries with at least one relevant D1 document and count only D1 documents as relevant; we proceed analogously for group D2. Note that this only touches on the calculation of the quality metrics over the retrieved answer list from the single mixed index I_T as a whole; all candidates remain in the list at their original ranks. Thus these are *not* searches of separate indexes and do not interfere with the execution of translators. The purpose is only to extract more accurate, fine-grained insight from evaluation results.

Retention of native retrieval metrics. Our main measure is nDCG@10, which rewards relevant answers near the top of the list. We also report MRR@10, the reciprocal rank of the first relevant answer within ten results, and Recall@100, the fraction of relevant answers found within one hundred. Write $M_G(I)$ for the mean query score of metric M on index I, evaluated for group G. We report

$$\mathrm{Ret}_{M,G}(T) = \frac{M_G(I_T)}{M_G(I_{\mathrm{native}})}, \qquad G \in \{\texttt{D1} \cup \texttt{D2}, \texttt{D1}, \texttt{D2}\}.$$

A value of 1 means equal retrieval quality, not identical rankings; values above 1 are possible because native rankings need not be optimal for relevance. We write R_1, R_2 and R for retention over D1, D2 and D1 ∪ D2, respectively. Their query sets and normalizers differ, so average of R_1 and R_2 is not R.



(a) Varying #anchors (|D0|)

(b) Anchors from outside the fit corpus (out-of-domain)

Figure 1: Impact of anchors over FEVER: (a) gte→kalm, random anchors, with #anchors (|D0|) from 64 to 1,048,576. (b) gte→kalm, results are averaged over ArguAna, FiQA, NFCorpus, and SciFact. Filled markers denote maps fitted and evaluated on the same corpus (in-domain); hollow maps fitted on FEVER (out-of-domain).

4 Evaluating Embedding Translation

We next evaluate translators with ETBench. We ask five questions. How do current translators perform? Does better alignment give better translation (RQ1)? What is the role of anchors (RQ2)? Does encoder geometry predict translation (RQ3)? Does alignment-translation distinction extend to alignment uses beyond mixed-index retrieval (RQ4)? What if multiple source encoders coexist (RQ5)?

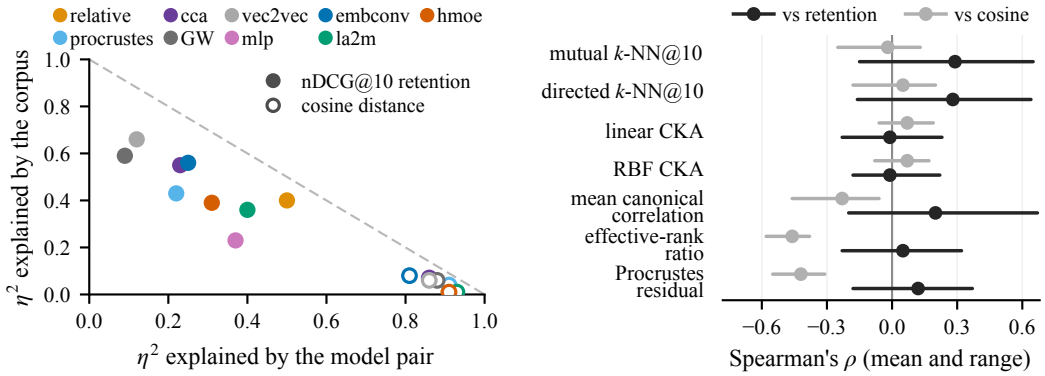
RQ1: How do current translators perform and does better alignment give better retrieval? Table 1 reports, for all 9 translators and `identity` control, mean alignment and retrieval results over the 24 encoder pairs and 4 corpora (see Appendix I for detailed results for every evaluation cell).

Alignment ≠ translation. The best alignment does not give the best translation, and held-out alignment and retrieval rank translators differently. `hmoe` retains the best full-index retrieval quality, nDCG@10 $R = 0.945$, yet its held-out `Hits@1`, 0.812, is seventh of the nine translators. `cca` has the highest held-out `Hits@1`, 0.936, and ranks fifth by translation retention ($R = 0.871$). `procrustes` is nearly the top for alignment, with `Hits@1` of 0.898, but gives R of only 0.603, even lower than `identity`, and as low as 0.293 for translated answers. Counterpart cosine of alignment is similar: `relative` has the highest held-out cosine, 0.978, but recovers only 0.520 of its counterparts and retains 0.796 of full nDCG@10. This is consistent with the observation in Section 2.2 that alignment counterpart cosine fixes the length of a translation error but not its direction, whereas retrieval depends on the projection of that error onto the query, measured against the ranking gap.

Aligning with target encodings is key. The three translation-oriented methods (Section 3) take the top three places in full-index nDCG@10 retention: `hmoe` (0.945), `la2m` (0.941) and `embconv` (0.883). Among alignment-oriented methods, `cca` is the strongest at 0.871, ahead of `relative` (0.796), while `procrustes` (0.603), `vec2vec` (0.599) and `GW` (0.593) fall below `identity` (0.664). Documents already encoded in the target space should therefore be exploited when fitting the alignment map, despite providing no anchor pairs to fit it directly.

In contrast, we find that similar source and target encoders do not always give better translation. Specifically, we grouped the encoder pairs into the same family at different sizes, a shared backbone, and different families, and tested translation quality of each translator. As shown in Table A32 of Appendix J, over `identity`, `la2m` gains 0.288 in nDCG@10 retention within an encoder family and 0.284 across families; similarly, `hmoe` gains 0.277 and 0.296, and `cca` 0.200 and 0.229.

Limitations of retrieval measures. Full-index scores hide losses on translated documents. `identity` inserts raw source vectors into the target space and still keeps $R = 0.664$, because its R_2 is 1.117 while its R_1 is 0.054. `vec2vec` shows the same pattern, $R = 0.599$ with $R_1 = 0.004$, and so, less sharply, does `procrustes`, $R_1 = 0.293$ against $R_2 = 0.842$. Native documents keep answering their queries, and an R_2 above 1 means they even gain ranks as translated competitors drop out of the top ten. The full score can thus stay moderate while translated documents are completely absent.



(a) Variance decomposition across pairs and corpora

(b) Geometry scores versus outcomes

Figure 2: Encoder geometry vs. mixed-index retrieval. (a) Shares of variation associated with encoder pairs and corpora on a 24×4 grid: hollow points show counterpart cosine; filled points show full-index retention. (b) Spearman correlations between geometry and alignment/translation performance, averaged over methods. Black denotes retention for translation and grey is cosine for alignment.

A translated-only score misleads as well. For instance, `relative` is best on `D1`, with nDCG@10 $R_1 = 1.001$, but loses on native answers with R_2 of 0.661 only; `hmoe` gives $(R_1, R_2) = (0.965, 0.929)$. Translation quality must therefore be read from the full score together with both group scores, since any one of the three alone can hide which documents lose over the translated mixed index.

RQ2: What are the impact of anchors? We next examine the role of anchors in `D0` on translation.

Benefiting from more anchors. We first see whether translators can benefit from more anchors. Figure 1a fixes `gte`→`kalm` and the held-out FEVER queries and grows a random anchor set from 64 to 1,048,576. Over the fourteen doublings, `procrustes` and `relative` gain about 0.0046 retention per doubling, while `mlp`, `embconv`, `la2m` and `hmoe` gain 0.0198–0.0307, at least 4.3 times as much. Those two lead at 64 anchors but are fifth and sixth at the largest budget; `procrustes` gains only 0.013 between 2,048 and 131,072 anchors. The largest budget covers 19.4% of FEVER.

Out-of-domain anchors. We further examine if out-of-domain anchors could be used for translation. To do so, we fit each method's alignment map on FEVER at its largest budget and test it on ArguAna, FiQA, NFCorpus and SciFact without refitting for translation. As shown in Fig. 1b, `procrustes` falls from 0.648 to a mean of 0.533 and `relative` from 0.772 to 0.675; the others lose 0.04–0.11. Overall, `hmoe`, `embconv`, `mlp` and `GW` have relatively smaller degradation compared to using in-domain anchors, in part due to their less rigid alignment maps.

Fitting vs. held-out alignment. Table 1 also shows why `D1` is withheld from fitting. On the anchors `D0`, `relative` reaches counterpart cosine 0.987 and `Hits@1` 0.997. On `D1` its cosine barely moves, to 0.978, but its `Hits@1` falls to 0.520: held-out translations stay close to their counterparts, yet about half land nearer another document. A map that fits the anchors well need not align held-out documents well, just as RQ1 shows that held-out alignment need not predict retrieval.

RQ3: Does geometry predict translation quality? Theorem 1 ties retrieval not to the error alone but to the error against each query's ranking gaps. We now ask whether the geometry of an encoder pair, which knows nothing about queries, can still anticipate retrieval.

To do so, we use variance decomposition. For one method and one outcome, either held-out counterpart cosine to the native target encoding or full-index `nDCG@10` retention, let y_{pc} be its value on encoder pair p and corpus c. The grid is complete over $P = 24$ pairs and $C = 4$ corpora (SciFact, NFCorpus, ArguAna, FiQA). With grand mean $\bar{y}$ and row and column means $\bar{y}_{p\cdot}$ and $\bar{y}_{\cdot c}$,

$$\eta^2_{\text{pair}} = \frac{C \sum_p (\bar{y}_{p\cdot} - \bar{y})^2}{\sum_{p,c} (y_{pc} - \bar{y})^2}, \qquad \eta^2_{\text{corpus}} = \frac{P \sum_c (\bar{y}_{\cdot c} - \bar{y})^2}{\sum_{p,c} (y_{pc} - \bar{y})^2}, \tag{3}$$

the share of the variance across the 96 cells that each factor (model pair and corpus) explains.

Table 2: KV-cache translation (Qwen3-4B→Qwen3-8B). Next-token agreement compares the translated and native target predictions at the first continuation position; depth 16 averages agreement through the first 16 positions. KL compares the two distributions. The last column reports MMLU Hendrycks et al. (2021) 5-shot accuracy, with retention against the native target (0.830).

	vector alignment		next-token outcome			downstream
mapper	Hits@1	cos	next-token agreement	depth 16	KL ↓	MMLU 5-shot acc. (ret.)
identity	.835	.688	.488	.545	.160	.250 (.301)
procrustes	.951	.827	.819	**.857**	.067	.750 (.904)
cca	.943	.842	.750	.773	.174	.713 (.859)
GW	**.982**	.873	.772	.781	.138	.263 (.317)
vec2vec	.041	.359	.056	.041	8.491	.267 (.322)
embconv	.971	.882	.549	.583	.163	.730 (.880)
la2m	.946	.849	**.869**	.807	**.058**	**.797 (.960)**
hmoe	.931	**.883**	.575	.673	.141	.263 (.317)

Data-oriented vs. *model-oriented*. We evaluate the impact of corpus and encoder pairs on alignment and translation, and find that the corpus matters more for translation than for alignment. For nine methods on a complete grid of 24 pairs and four corpora, we split the variation of each outcome into the share associated with the encoder pair and the share associated with the corpus (Figure 2a). For held-out counterpart cosine, *i.e.,* alignment quality, the pair accounts for 0.81–0.93 of the variation and the corpus for 0.01–0.08, *i.e.,* alignment is largely a property of the pair. For nDCG@10 retention over the mixed index, *i.e.,* translation quality, the pair share drops to 0.09–0.50 and the corpus share rises to 0.36–0.66. That is, the same map, with nearly the same alignment, retrieves differently on different corpora. This suggests that, unlike alignment, translation is also data-dependent. As a result, the performance of translators cannot be simply estimated by the quality of the alignment map it uses.

Encoder geometry vs. *translation quality*. We examine if geometric properties of encoder pairs would imply translation quality. We compute seven scores of how the two encoders agree on D1, comparing nearest neighbors, similarity structure, linear fits, and the spread of the vectors across directions, over 96 evaluation cells, and correlate each score's ranking of the cells with each method's retention ranking (Spearman correlation, Figure 2b). The best mean correlation is $\rho = 0.29$, for how often the two encoders give a document the same ten nearest neighbors ("mutual k-NN@10"), but it ranges from -0.15 to 0.65 across methods, and every score is positive for some methods and negative for others. Even the fit by a rotation and a scale ("Procrustes residual"), mean 0.12, turns negative for one method. This shows that geometric properties of the encoders cannot robustly predict translation.

RQ4: Does the distinction extend beyond mixed-index retrieval? We next examine if the subtleties around alignment extend to uses of alignment maps and translated vectors beyond retrieval. We test two cases over translated vectors: continuing text from a translated attention key–value (KV) cache for LLMs, and mixed-population clustering for translated documents along with native ones.

KV-cache translation. We map caches from generative language model Qwen3-4B to Qwen3-8B (not Qwen3 embedding encoders of RQ1), install the translated or the native cache in the target, and feed both the same true continuation tokens (Appendix F). Table 2 reports how often the two caches give the same next-token prediction at the first scored position and on average over sixteen positions, and the Kullback–Leibler (KL) divergence between their output distributions (lower being closer). la2m raises first-position agreement from the control's 0.488 to 0.869 with the lowest KL, 0.058; procrustes, weak in retrieval, leads over sixteen positions at 0.857; hmoe, the best retriever, reaches only 0.575 at the first position; vec2vec falls to 0.056. On 5-shot MMLU (Hendrycks et al., 2021) benchmark, la2m performs best at 0.797, while hmoe, GW, and vec2vec remain near chance floor (0.25). Notably, GW achieves the highest cache Hits@1 (0.982) but only 0.263 score, showing that counterpart recovery does not guarantee downstream utility.

Mixed-population clustering. We further tested the eight clustering tasks from the MTEB benchmark Muennighoff et al. (2023). We fit cluster centers on native D2 and ask whether each translated D1 vector chooses the same center as its native counterpart; we also examine retention of adjusted mutual information (AMI) and centroid agreement (see Table A7 in Appendix H). Six fitted maps reach

Table 3: Two sources translated into one index, averaged over three directions (nv-embed-v2 ← gritlm, gte; qwen3-8b ← qwen3-0.6b, gritlm; gte ← openai, linq) and four corpora (arguana, fiqa, nfcorpus, scifact). Each entry is nDCG@10 retention for results with a full translation, but calculated per group setting. For D1 group: (a) and (b) are the halves translated from the first and the second source, and *mixed* pools them. For D2 group, since all D2 documents are native in the target space, we have only one retention for each translator, as the quality of its full-index retrieved results projected to D2. For full group, Δ is the overall retention change against the better of the two sources translated alone. Rank is by full group retention. Results are averaged over translation directions and corpora; see Appendix K (Table A33) for details.

	D1				full		
mapper	(a)	(b)	mixed	D2	ret.	Δ	rank
identity	.085	.000	.043	1.131	.661	↓.189	–
relative	**.967**	.854	**.911**	.832	.862	↓.053	5
procrustes	.209	.169	.189	1.056	.680	↓.169	8
cca	.714	.835	.774	.990	.893	↓.055	4
GW	.234	.159	.196	1.113	.718	↓.173	7
mlp	.693	.760	.726	1.013	.896	↓.053	3
vec2vec	.000	.000	.000	1.128	.639	↓.194	–
embconv	.626	.630	.628	1.042	.861	↓.081	6
la2m	.708	.687	.698	1.066	.914	↓.054	2
hmoe	.816	**.887**	.852	1.012	**.945**	↓.032	1

Hits@1 of 0.917–0.992, yet cluster very differently: procrustes and cca (Hits@1 0.990 and 0.991) retain 0.964 and 0.445 of AMI, while la2m and hmoe, lowest in Hits@1 (0.968 and 0.917), cluster best (AMI retention 0.999 and 0.966).

For both cases above, we have observed a similar pattern as shown in RQ1: when translated vectors are used for downstream tasks other than alignment counterpart recovery, usual alignment quality does not apply to the performance of functions that use the translated vectors.

RQ5: What changes with multiple sources? We finally return to retrieval with embeddings from two source encoders, translated into one target index. We split D1 into two halves, each assigned to a distinct source encoder; we translate each with a map fitted on its source's anchor pairs, and store both beside native D2. Table 3 reports means over 3 translation directions and 4 corpora.

The results broadly resemble single-source cases in RQ1. For instance, hmoe and la2m again lead in mean full-index nDCG@10 retention, while GW, procrustes, and vec2vec remain at the bottom. Moreover, each translator's performance degrades compared to its best single-source translation. Specifically, all methods' Δ values are negative. hmoe has the smallest mean shortfall, 0.032, and the highest full retention, $R = 0.945$; procrustes and GW trail their respective references by 0.169 and 0.173. This comparison is to the better source, so it does not isolate the effect of source count.

The same applies to quality measures. RQ1 suggests that an aggregated overall score would hide translation losses, and it is similar for multi-source translation. For instance, identity retains 0.661 overall but only 0.043 on pooled D1, while native D2 reaches 1.131. Conversely, relative leads on pooled D1 (0.911) but has the lowest D2 retention (0.832). Even the two translated halves differ: for the qwen3-8b target, la2m retains 0.774 on qwen3-0.6b documents and 0.518 on gritlm documents, averaged over corpora (Appendix K). Thus the lesson of RQ1 extends to multiple sources, which is to evaluate each source and the native documents alongside the full index.

5 Conclusion

While embedding translation uses alignment maps to translate vectors across embedding spaces, in contrast to common belief, the quality of translation is not accurately reflected by the quality of alignment. We characterize the connection between alignment and translation formally, and develop a tight condition on when alignment can certify translation results. We further develop ETBench, a benchmark for embedding translation. The results expose translated-answer losses hidden by alignment success or aggregate utility measures, and show that tailored alignment with native embeddings in the target space can help with overall translation quality.

USE OF LARGE LANGUAGE MODELS

We disclose our use of LLMs by task, following the ICLR 2027 policy on AI use. We used LLM agents to help write code, improve the manuscript, set up and run experiments in parallel on our computing resources, write scripts to collect and combine results automatically, and manage the experiments.

REPRODUCIBILITY STATEMENT

We provide scripts to reproduce all reported results in our anonymous repository (`https://anonymous.4open.science/r/ETBench-321C/`). We regret that we are unable to release the approximately 3 TB embedding dataset during anonymous review. However, the models and source datasets we used are publicly available, subject to any access requirements set by their providers. Researchers can use these resources and our scripts to generate the embeddings and reproduce the results. We will release the embedding dataset once the anonymous review period ends.

REFERENCES

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1881–1890, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1214. URL https://aclanthology.org/D18-1214/.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. Llama-Embed-Nemotron-8B: A universal text embedding model for multilingual and cross-lingual tasks, 2025. URL https://arxiv.org/abs/2511.07025.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.137.

Chanyeol Choi, Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, and Jy-yong Sohn. Linq-Embed-Mistral technical report, 2024. URL https://arxiv.org/abs/2412.03223.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 462–471, Gothenburg, Sweden, 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1049. URL https://aclanthology.org/E14-1049/.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 710–721, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1070. URL https://aclanthology.org/P19-1070/.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3887–3896, 2020.

Sid Ali Hamideche, Louis-Adrien Dufrène, Quentin Lampin, and Guillaume Larue. How far do simple transformations translate across text embedding models?, 2026. URL https://arxiv.org/abs/2608.05980.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Taekyung Heo, Rasoul Shafipour, Ritchie Zhao, Maximilian Golub, Mohammad Mahdi Kamani, Ritika Borkar, Makesh Tarun Chandran, Pantea Zardoshti, and Bita Darvish Rouhani. Cross-model KV cache transfer in LLM families: A closed-form linear mapping for prefill reuse, 2026. URL https://arxiv.org/abs/2608.03893.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Florian Jaeckle, Fartash Faghri, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Fastfill: Efficient compatible model update. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rnRiiHw8Vy.

Rishi Dev Jha, Collin Zhang, Vitaly Shmatikov, and John Xavier Morris. Harnessing the universal geometry of embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=jiCLUPq5xv.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uDjc7tTWmb.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023. URL https://arxiv.org/abs/2308.03281.

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

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. SFR-Embedding-Mistral: Enhance text retrieval with transfer learning. Salesforce AI Research Blog, 2024. URL https://www.salesforce.com/blog/sfr-embedding/.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation, 2013. URL https://arxiv.org/abs/1309.4168.

Mistral AI. La plateforme. Mistral AI Blog, December 2023. URL https://mistral.ai/news/la-plateforme.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SrC-nwieGJ.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BC4lIvfSzv.

OpenAI. New embedding models and API updates. OpenAI Blog, January 2024. URL https://openai.com/index/new-embedding-models-and-api-updates/.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2664–2672. PMLR, 2016.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5919–5930. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.477. URL https://aclanthology.org/2020.emnlp-main.477/.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019. URL https://arxiv.org/abs/1706.04902.

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451.

Erich Schubert. A triangle inequality for cosine similarity. In *Similarity Search and Applications (SISAP)*, volume 13058 of *Lecture Notes in Computer Science*, pp. 32–44. Springer, 2021. doi: 10.1007/978-3-030-89657-7_3.

Seonguk Seo, Mustafa Gokhan Uzunbas, Bohyung Han, Sara Cao, and Ser-Nam Lim. Metric compatible training for online backfilling in large-scale retrieval, 2023. URL https://arxiv.org/abs/2301.03767.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6367–6376, 2020. doi: 10.1109/CVPR42600.2020.00640.

Mohamed A. Soliman and Ihab F. Ilyas. Ranking with uncertain scores. In *IEEE International Conference on Data Engineering (ICDE)*, pp. 317–328, 2009. doi: 10.1109/ICDE.2009.102.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2022. URL https://arxiv.org/abs/2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 text embeddings: A technical report, 2024. URL https://arxiv.org/abs/2402.05672.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649. ACM, 2024. doi: 10.1145/3626772.3657878.

Beining Yang and Yang Cao. Generalizable and composable multi-model embedding translation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=qmfp2eqYD1.

Beining Yang, Yang Cao, and Yang Ren. Integrating vector databases across embedding models. *Proc. ACM Manag. Data*, 3(6), December 2025. doi: 10.1145/3769803. URL https://doi.org/10.1145/3769803.

Jiayi Yao, Hanchen Li, Yuhan Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. CacheBlend: Fast large language model serving for RAG with cached knowledge fusion. In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys '25)*, pp. 94–109. ACM, 2025. doi: 10.1145/3689031.3696098.

Jinsung Yoon and Sercan O Arik. Embedding-converter: A unified framework for cross-model embedding transformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25464–25482, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1237. URL https://aclanthology.org/2025.acl-long.1237/.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. KaLM-Embedding-V2: Superior training techniques and data inspire a versatile embedding model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Y7qzhvWhcz.

Appendix

A Proofs and extensions of Theorem 1 14

B Related Work 18

C Encoders 19

D Datasets 20

E Translation methods 21

F KV-cache setup 22

G Efficiency Comparison 23

H Mixed-population clustering 23

I The per-pair results of Table 1 23

J Results by encoder relationship 49

K The per-group results of Table 3 49

A Proofs and extensions of Theorem 1

We prove Theorem 1 and discuss its extensions to support ordering-based retrieval metrics like nDCG.

Setting and notation. We start with the notations. Following Section 2, `D0` is the set of anchors that serve only to fit the map; the candidates are $\texttt{D1} \cup \texttt{D2}$, with $N = |\texttt{D1} \cup \texttt{D2}|$. For a candidate d, $z_d = f_t(d)$ is its native target vector, which exists on `D1` in the hypothetical all-native index I_{native} as reference, even though the translator never sees it, and v_d is the vector the mixed index I_T stores: $T(f_s(d))$ on `D1` and z_d on `D2`. The all-native index I_{native} stores z_d for every candidate. We fix a unit target query $q \in \mathbb{R}^{d_t}$, $d_t \geq 2$, take every z_d and v_d unit, and rank by decreasing $q^\top v$. The native angle $\theta_d = \arccos(q^\top z_d) \in [0, \pi]$ is the query's view of candidate d: ranking q's candidate answers by score is ranking by their θ_d *w.r.t.* q in increasing order. We write $R_k = q(I_{\text{native}})$ for the native top-k set, $1 \leq k < N$, and assume its boundary is strict, $\theta_d < \theta_b$ whenever $d \in R_k$ and $b \notin R_k$.

For convenience, Table A1 collects notations used in this appendix.

An α_d-bounded map may place v_d anywhere in the *spherical cap* of angular radius α_d about z_d,

$$C_d = \{v \in \mathbb{R}^{d_t} : \|v\|_2 = 1,\ \arccos(z_d^\top v) \leq \alpha_d\}, \qquad 0 \leq \alpha_d \leq \pi.$$

For unit vectors, ranking scores are cosines. Hence a query sees angles when ranking candidate documents. And on the unit sphere the cap contains exactly the outputs whose error $e_d = v_d - z_d$ has length $\|e_d\|_2 \leq 2\sin(\alpha_d/2)$, so for one document an angular bound and a length bound carry the same information (Section 2.2).

Remark. Theorem 1(a) speaks of every α_d-bounded map, and its necessity half must exhibit a map that occupies a chosen point of every cap at once. We therefore allow arbitrary maps with unit outputs and assume that the source vectors in `D1` are pairwise distinct. Any assignment of one output to each of these inputs, together with an arbitrary unit vector on every other input, is a function $T : \mathbb{R}^{d_s} \to \mathbb{R}^{d_t}$, so every point of $\prod_d C_d$ is realized by one map. Theorem 1(b) shows that the failure the bounds permit already occurs inside the orthogonal family. With several source encoders nothing changes: each map contributes its own translated block, and we impose the distinct-input assumption per map.

We start with a lemma about angular bounds that we will use for proving Theorem 1(a).

Lemma A.1 *Let $q, z \in \mathbb{R}^p$ be unit vectors, $p \geq 2$, and let $\theta = \arccos(q^\top z)$. For $0 \leq \alpha \leq \pi$, the*

Table A1: Notation of this appendix. The first block is shared with Section 2; the second belongs to the construction of part (b).

symbol	meaning
z_d, v_d	native target vector of candidate d, and the vector the mixed index stores: $T(f_s(d))$ on `D1`, z_d on `D2`
$e_d = v_d - z_d$	translation error; zero on `D2`
$\theta_d = \arccos(q^\top z_d)$	native angle between the query and d; a smaller angle is a higher native score
α_d, C_d	angular error bound, and the cap of unit vectors within angle α_d of z_d; $\alpha_d = 0$ on `D2`
$[L_d, U_d]$	the scores $q^\top v$ that the cap C_d allows (Lemma A.1); $L_d = U_d = q^\top z_d$ on `D2`
R_k, $\widehat{R}_k$	native and served top-k sets of the query
H, G	orthogonal planes of $\mathbb{R}^4$: H carries the queries and the scores, G carries document identities and the anchors
c, r	shares of a unit document vector in H and G, $c^2 + r^2 = 1$
ϕ	native angle from the queries to the relevant documents
σ, η	rotations of H under the failing map T and the succeeding map T'; $\sigma \le \eta$
$\beta(t)$	angular error on the sphere produced by rotating the H-component by t
t_j, ξ_i	small offsets that make the document and query orders strict

scores $q^\top v$ over unit vectors satisfying $\arccos(z^\top v) \le \alpha$ form the interval

$$[L, U], \qquad L = \cos(\min\{\pi, \theta + \alpha\}), \quad U = \cos(\max\{0, \theta - \alpha\}). \tag{4}$$

Every value in this interval is attained in a plane containing q and z.

Proof. The proof follows from the triangle inequality for cosine similarity over spherical geometry (Schubert, 2021): we describe q and v relative to z, and the extreme scores occur when v moves in the plane of q and z, toward or away from q. For $0 < \theta < \pi$, write $q = \cos\theta\, z + \sin\theta\, h$, where h is unit and perpendicular to z. Every allowed vector has the form $v = \cos\delta\, z + \sin\delta\, g$, with $0 \le \delta \le \alpha$ and g unit and perpendicular to z; when $\sin\delta = 0$, any such g suffices. Thus

$$q^\top v = \cos\theta\cos\delta + \sin\theta\sin\delta\, h^\top g, \qquad -1 \le h^\top g \le 1.$$

For fixed δ, we attain the largest score $\cos(\theta - \delta)$ at $g = h$ and the smallest $\cos(\theta + \delta)$ at $g = -h$. Optimizing over $\delta \in [0, \alpha]$ gives U at $\delta = \min\{\alpha, \theta\}$ and L at $\delta = \min\{\alpha, \pi - \theta\}$: moving toward q helps until v reaches q, and moving away hurts until v reaches $-q$, so the smallest and largest angles between q and v are $\theta - \min\{\alpha, \theta\} = \max\{0, \theta - \alpha\}$ and $\theta + \min\{\alpha, \pi - \theta\} = \min\{\pi, \theta + \alpha\}$. This is where the $\max$ and $\min$ in Equation 4 come from: they hold both angles inside $[0, \pi]$, because v can come no closer to q than q itself and no farther from it than $-q$. If $q = z$, the range is $[\cos\alpha, 1]$; if $q = -z$, it is $[-1, -\cos\alpha]$, as claimed. In these two cases choose any unit $h \perp z$, which exists since $p \ge 2$. In every case the path $v(t) = \cos t\, z + \sin t\, h$, $-\alpha \le t \le \alpha$, stays in the cap and passes through both extremizers. Continuity supplies every intermediate score. □

Proof of Theorem 1(a). The proof follows the usual separation analysis of interval-valued scores (Soliman & Ilyas, 2009), where in our case the intervals are derived from angular errors. More specifically, we use Lemma A.1 to turn each candidate into an interval. With $z = z_d$ and $\alpha = \alpha_d$, let $[L_d, U_d]$ be the range of $q^\top v_d$ over the cap C_d: L_d is the score after the worst push away from the query, U_d the score after the best pull toward it. A native candidate has $L_d = U_d = q^\top z_d$. Theorem 1(a) now asks whether every member's interval lies strictly above every nonmember's.

For a boundary pair $d \in R_k$, $b \notin R_k$, cosine is strictly decreasing on $[0, \pi]$, so

$$L_d > U_b \iff \min\{\pi, \theta_d + \alpha_d\} < \max\{0, \theta_b - \alpha_b\} \iff \alpha_d + \alpha_b < \theta_b - \theta_d.$$

The second equivalence drops the $\min\{\pi, \cdot\}$ and $\max\{0, \cdot\}$ that Lemma A.1 places around the angles. This is because an angle between two directions never exceeds π and never falls below 0: once the worst push has carried v_d onto the direction opposite to q, the score has reached -1 and a larger bound cannot lower it further; once the best pull has carried v_b onto q itself, the score has reached 1 and cannot rise further. The two numbers compared in the middle of the display are therefore the member's largest possible angle and the nonmember's smallest possible angle, each held inside $[0, \pi]$.

Suppose that inequality holds. Its right side is at most $\theta_b \le \pi$, so its left side is strictly below π: the $\min$ did nothing and the left side is $\theta_d + \alpha_d$ itself. Its left side is at least 0, so its right side is strictly above 0: the $\max$ did nothing and the right side is $\theta_b - \alpha_b$ itself. Hence the inequality holds without them. Conversely, if $\theta_d + \alpha_d < \theta_b - \alpha_b$, the smaller of the two is at least 0 and the larger is at most $\theta_b \le \pi$, so both already lie in $[0, \pi]$ and the $\min$ and $\max$ leave them unchanged. Taking the worst pair, Equation (2) of Theorem 1(a) is equivalent to

$$\min_{d \in R_k} L_d > \max_{b \notin R_k} U_b. \tag{5}$$

Sufficiency. Every allowed output gives a score in $[L_d, U_d]$. Under Equation (5), each member of R_k strictly outscores each nonmember, so every bounded map has R_k as its unique top-k set.

Necessity. The witness is the obvious one: push the most exposed member to the bottom of its interval and pull the strongest nonmember to the top of its. If the condition fails, choose $d_0 \in R_k$ minimizing L_d and $b_0 \notin R_k$ maximizing U_b, so $L_{d_0} \le U_{b_0}$. We place these two candidates at their score extremizers and leave every other candidate at its native vector. Lemma A.1 and the product-of-caps freedom realize this configuration by one bounded map. At least one of the two is translated: if both were native, the strict boundary would give $L_{d_0} = q^\top z_{d_0} > q^\top z_{b_0} = U_{b_0}$. If $L_{d_0} < U_{b_0}$, a nonmember outranks a member, so R_k is not the returned top-k set. At equality, every other member scores at least this value and every other nonmember at most it, so exchanging d_0 and b_0 gives a second top-k set: uniqueness fails even if a tie-breaking rule happens to return R_k. □

Proof of Theorem 1(b). Fix $\varepsilon > 0$, $\rho \in (0, 1]$, and integers $n \ge 0$, $m, k \ge 1$. We construct n anchors and m queries in four dimensions, which proves the result for any finite query count as well as the single fixed query case stated in Theorem 1(b).

Overall idea. The construction has to do three things at once: keep both maps orthogonal and exact on every anchor, so that no anchor statistic tells them apart; keep every held-out counterpart recoverable with errors as small as we like, so that held-out `Hits@1` does not distinguish them; and still make them disagree about mixed retrieval. We do this with two orthogonal planes. The *identity plane* G carries each document's identity and all the anchors; the *query plane* H carries the queries and hence the scores. Each document has a component in both. A map that rotates H and fixes G moves every translated score by the same angle while leaving identities and anchors untouched, so it is invisible to anchor fit and to counterpart matching but visible to every query. The size of the rotation sets the alignment error; its sign decides whether translated documents move toward or away from the query, and therefore whether they keep or lose their place against the native documents, which do not move.

Construction. Split $\mathbb{R}^4$ into orthogonal planes H and G, and write $u(t) = (\cos t, \sin t)$ in coordinates of H. Choose distinct unit vectors $h_1, \dots, h_k$ in G, one identity per relevant document. For $k \ge 2$, let $\Delta = \min_{j \ne \ell} \|h_j - h_\ell\|_2 > 0$; for $k = 1$, set $\Delta = 1$. We take n distinct unit vectors of G as the anchors and give every document identical source and target encodings, so that the identity map would be a perfect translator and the anchors reveal nothing about H. Set $c = 1/2$ and $r = \sqrt{1 - c^2}$, the shares of each unit document vector in H and G, and $\phi = \pi/6$, the native angle at which the relevant documents sit to one side of the queries. Rotating a document's H-component by t produces the angular error

$$\beta(t) = 2 \arcsin(c \sin(t/2)), \qquad 0 \le t \le \pi/12,$$

on the sphere, smaller than t because only the H-share moves. Choose

$$0 < \eta < \min\{\pi/12,\ \varepsilon,\ r\Delta/(2c)\}, \qquad \beta(\sigma) = \rho\, \beta(\eta), \quad 0 < \sigma \le \eta.$$

Here η is the rotation of the map that will succeed and σ the rotation of the map that will fail; the second equation fixes their error ratio at ρ. The choice of σ is unique because β is continuous and strictly increasing on this interval with $\beta(0) = 0$, and $\sigma = \eta$ when $\rho = 1$. For $t_j = j\sigma/(32(k+1)) \in (0, \sigma/32)$, small offsets that make every order strict, define

$$z_{a_j} = c\, u(\phi + t_j) + r\, h_j, \qquad z_{b_j} = c\, u(-\phi - \sigma/2 - t_j) + r\, h_j \qquad (1 \le j \le k).$$

The a_j are the relevant documents, placed in `D1`; the b_j are irrelevant native documents, placed in `D2` on the other side of the queries and $\sigma/2$ further out, so that the native gap between the two blocks is of order σ. This gives $N = 2k$ unit candidates; the anchors remain outside the index. Choose distinct $\xi_i \in (-\sigma/32, \sigma/32)$ and let $q_i = u(\xi_i)$, $1 \le i \le m$: the queries lie in H, close to its first axis. For

every query, assign relevance one to each a_j and zero to each b_j. Let T rotate H by $+\sigma$, away from the queries, and T' rotate H by $-\eta$, toward them, both fixing G pointwise. Both maps are orthogonal; we apply them only to `D1`.

We now check if the construction satisfies the three criteria described at the beginning, *i.e.,* (1) whether both maps fit every anchor exactly, (2) whether both maps have identical, perfect alignment accuracy, measured by their held-out counterpart recovery over `D1`, and (3) whether they produce mixed indices with contrasting retrieval results.

(1) anchor fit and held-out angular error. Both maps fix every anchor, since the anchors lie in G, and preserve all distances within the translated block, since they are rigid. For either signed rotation $t \in \{\sigma, -\eta\}$, the image of z_{a_j} is $w_j(t) = c\,u(\phi + t_j + t) + r\,h_j$, and

$$z_{a_j}^\top w_j(t) = r^2 + c^2 \cos t, \qquad \|w_j(t) - z_{a_j}\|_2 = 2c|\sin(t/2)|.$$

The angular error is therefore $\beta(|t|)$, the same for every held-out document. We obtain $\beta_T = \beta(\sigma) = \rho\,\beta(\eta) = \rho\,\beta_{T'}$. Since $0 < c < 1$, we have $0 < \beta(t) < t$ for $0 < t \le \eta$, hence $0 < \beta_T \le \beta_{T'} < \eta < \varepsilon$.

(2) Counterpart recovery (alignment error on `D1`). Counterpart recovery is decided in G, which the maps do not touch. For either map, let w_j be the image of z_{a_j}. Its distance to its own counterpart is at most $2c\sin(\eta/2) < c\eta$. For $\ell \neq j$, the G-component of $w_j - z_{a_\ell}$ has length at least $r\Delta > 2c\eta$. Thus z_{a_j} is strictly nearer than every other vector in $f_t(\texttt{D1})$, and both maps have perfect held-out `Hits@1`. This remains true if the counterpart pool also includes `D2` and the anchors: the angle between the H-components of w_j and z_{b_ℓ} is $2\phi + 3\sigma/2 + t_j + t_\ell$ under T and $2\phi + \sigma/2 - \eta + t_j + t_\ell$ under T'. Both lie in (ϕ, π), since $\sigma \le \eta < \pi/12$ and $0 < t_j, t_\ell < \sigma/32$. Their distance in H alone exceeds $2c\sin(\phi/2) > c\eta$, using $2\sin(\pi/12) > \pi/12 > \eta$. An anchor has zero H-component, so its distance from w_j is at least $c > c\eta$.

(3) Mixed index retrieval result. Retrieval is decided in H, which the maps rotate. The native scores $s(q,d) = q^\top z_d$ and mixed scores $\widehat{s}_T(q,d) = q^\top v_d$ are

$$\begin{aligned} s(q_i, a_j) &= c\cos(\phi + t_j - \xi_i), & s(q_i, b_j) &= c\cos(\phi + \sigma/2 + t_j + \xi_i), \\ \widehat{s}_T(q_i, a_j) &= c\cos(\phi + \sigma + t_j - \xi_i), & \widehat{s}_{T'}(q_i, a_j) &= c\cos(\phi - \eta + t_j - \xi_i). \end{aligned}$$

All four cosine arguments lie in $(0, \pi/2)$, where cosine decreases, so a smaller argument is a higher score. Natively, the a_j arguments are below $\phi + 2\sigma/32$, while every b_j argument lies between $\phi + 15\sigma/32$ and $\phi + 18\sigma/32$: all relevant documents precede all irrelevant ones, with a gap of order σ. Under T, every a_j argument exceeds $\phi + 31\sigma/32$, so the rotation by σ carries the whole relevant block past the irrelevant one. Under T', every a_j argument is below $\phi - \eta + 2\sigma/32 \le \phi - 30\sigma/32$, so the relevant block moves further ahead. Within each block, scores strictly decrease with t_j in all three indexes. Consequently T' preserves the entire native order, while T exchanges the two blocks without changing their internal orders. In particular, for every q_i,

$$q_i(I_T) = \{b_1, \ldots, b_k\}, \qquad q_i(I_{T'}) = q_i(I_{\text{native}}) = \{a_1, \ldots, a_k\},$$

with unique top-k sets (hence, nDCG@k, MRR@k, and Recall@k are 1 for I_{native} and $I_{T'}$, and 0 for I_T). Together with Steps (1)–(2) above, this proves Theorem 1(b). □

Remark (anchor span). The construction places every anchor in G, so the anchors leave H unconstrained; this is what lets two orthogonal maps fit every anchor exactly yet differ on the held-out documents. If the anchor source vectors spanned the space, exact fit would determine a linear map uniquely, here the identity, and no such pair would exist. Part (b) therefore shows that anchor fit and held-out alignment scores do not certify retrieval; it does not describe a map fitted to spanning anchors. Real anchors typically span the space but are fitted only approximately, and the residual then carries the same freedom, which part (a) bounds through α_d.

From Recall@k to nDCG@k. Section 2.3 states that preserving the top-k set preserves Recall@k, and that preserving its internal order as well gives identical nDCG@k. The same intervals give the exact condition: to protect the list, we must protect its boundary and every adjacent pair inside it.

Suppose the native scores within the exact top-k answer set R_k are distinct, and write its native order as $\ell = (d_1, \ldots, d_k)$. Then we have the following:

Proposition A.1 (nDCG) *Every bounded alignment map returns the same ordered answer list ℓ with strict scores inside it and across its boundary if and only if Equation (5) holds and*

$$L_{d_i} > U_{d_{i+1}} \qquad (1 \le i < k). \tag{6}$$

Equivalently, Equation (2) holds and $\alpha_{d_i} + \alpha_{d_{i+1}} < \theta_{d_{i+1}} - \theta_{d_i}$ for every $i < k$.

Proof. Under these conditions, each adjacent pair satisfies $q^\top v_{d_i} \ge L_{d_i} > U_{d_{i+1}} \ge q^\top v_{d_{i+1}}$ under every map, and Equation (5) excludes all nonmembers. Conversely, a failed boundary condition has the witness from part (a). If an adjacent condition fails, we place d_i at L_{d_i} and d_{i+1} at $U_{d_{i+1}}$, leaving all other vectors native; at least one of the pair is translated, since native scores within R_k are distinct. The cap extremizers and distinct source inputs realize a bounded map with a tie or reversal in this pair. The angular form follows from the equivalence $L_d > U_b \iff \alpha_d + \alpha_b < \theta_b - \theta_d$ established in the proof of Theorem 1(a): Equation (6) becomes $\alpha_{d_i} + \alpha_{d_{i+1}} < \theta_{d_{i+1}} - \theta_{d_i}$, and Equation (5) is Equation (2) by the same proof. □

Characterization with translated mixed index I_T. Theorem 1(a) uses the native angles as well as the error bounds, which may not be helpful if one only holds the translated mixed index I_T, as is usual in practice. Below we derive a variant of characterization between alignment and translation in Theorem 1(a) using certifying machinery from only the translated mixed index I_T, *i.e.,* served vectors v_d in I_T and valid bounds α_d that one would like to consider. The idea is similar: we fix v_d, let the unknown native vectors vary in the caps about them *w.r.t.* α_d that we care, and ask whether every consistent native index has the served answer. Write $\widehat{\theta}_d = \arccos(q^\top v_d)$ and let $\widehat{R}_k = q(I_T)$ have a strict served boundary (we do not assume a strict native boundary here).

Proposition A.2 *Fix the unit query and served vectors. Suppose each unknown unit z_d on `D1` may vary independently subject only to $\arccos(v_d^\top z_d) \le \alpha_d$, while $z_d = v_d$ and $\alpha_d = 0$ on `D2`. Every consistent native index has $\widehat{R}_k$ as its unique top-k set if and only if*

$$\alpha_d + \alpha_b < \widehat{\theta}_b - \widehat{\theta}_d \quad \textit{for every } d \in \widehat{R}_k,\ b \notin \widehat{R}_k. \tag{7}$$

Proof. By Lemma A.1, each native score ranges over $[\widehat{L}_d, \widehat{U}_d]$, where $\widehat{L}_d = \cos(\min\{\pi, \widehat{\theta}_d + \alpha_d\})$ and $\widehat{U}_d = \cos(\max\{0, \widehat{\theta}_d - \alpha_d\})$. Every configuration of these scores is attainable under the stated independent-cap assumption. If $\min_{d \in \widehat{R}_k} \widehat{L}_d > \max_{b \notin \widehat{R}_k} \widehat{U}_b$, every member strictly outscores every nonmember natively. Otherwise, choose a member attaining the smallest lower endpoint and a nonmember attaining the largest upper endpoint, leaving all other native vectors equal to their served vectors. This permits a native boundary tie or reversal and defeats the universal unique-set guarantee. It remains to see that the interval condition is Equation (7). For each pair, $\widehat{L}_d > \widehat{U}_b$ holds exactly when $\alpha_d + \alpha_b < \widehat{\theta}_b - \widehat{\theta}_d$, by the step in the proof of Theorem 1(a) that removes the $\min$ and $\max$ around the angles: that step used only that the angles lie in $[0, \pi]$ and that cosine decreases there, and both remain true with the served angles $\widehat{\theta}_d$ in place of the native ones. □

B Related Work

Alignment benchmarks. Bilingual lexicon induction (BLI) evaluates whether a mapped word vector retrieves its translation from a target vocabulary. MUSE provides a common alignment and evaluation pipeline (Lample et al., 2018), while Glavaš et al. (2019) show that BLI performance need not predict downstream cross-lingual retrieval quality. ETBench's held-out `Hits@1` likewise measures correspondence, using the same document encoded by the target model as the positive. Its primary evaluation, however, concerns a different ranking problem: whether translated document vectors preserve the target model's retrieval results when they compete with native target vectors in a mixed index. ETBench compares this index with a fully native target index and reports retrieval quality separately for translated and native documents.

Retrieval benchmarks. BEIR provides query–document relevance judgments across diverse domains (Thakur et al., 2021), while MTEB evaluates embedding models on retrieval and other tasks (Muennighoff et al., 2023). ETBench reuses BEIR's corpora and judgments to evaluate a different

Table A2: The 18 encoders. *Params* are counted from the released weights. *Used*: M = main grid (Table A3), S = anchor ladder (Section 4 (RQ2)), C = clustering (Section 4 (RQ4)); — = in the release but in no experiment reported here. `linq` and `sfr` are both fine-tuned from `e5-mistral-7b-instruct`.

alias	checkpoint	backbone	params	dim	used
`bge-small`	`BAAI/bge-small-en-v1.5` (Xiao et al., 2024)	BERT	33M	384	M
`bge-large`	`BAAI/bge-large-en-v1.5` (Xiao et al., 2024)	BERT	335M	1024	M
`bge-m3`	`BAAI/bge-m3` (Chen et al., 2024)	XLM-R	568M	1024	M
`gte-small`	`thenlper/gte-small` (Li et al., 2023)	BERT	33M	384	M
`gte-large`	`thenlper/gte-large` (Li et al., 2023)	BERT	335M	1024	M
`gte`	`Alibaba-NLP/gte-Qwen2-7B-instruct` (Li et al., 2023)	Qwen2	7.6B	3584	M, S, C
`e5-large-v2`	`intfloat/e5-large-v2` (Wang et al., 2022)	BERT	335M	1024	M
`me5-large`	`intfloat/multilingual-e5-large-instruct` (Wang et al., 2024)	XLM-R	560M	1024	M
`qwen3-0.6b`	`Qwen/Qwen3-Embedding-0.6B` (Zhang et al., 2025)	Qwen3	0.6B	1024	M
`qwen3-8b`	`Qwen/Qwen3-Embedding-8B` (Zhang et al., 2025)	Qwen3	7.6B	4096	M
`gritlm`	`GritLM/GritLM-7B` (Muennighoff et al., 2025)	Mistral-7B	7.2B	4096	M, C
`nv-embed-v2`	`nvidia/NV-Embed-v2` (Lee et al., 2025)	Mistral-7B	7.9B	4096	M, C
`linq`	`Linq-AI-Research/Linq-Embed-Mistral` (Choi et al., 2024)	Mistral-7B	7.1B	4096	M, C
`sfr`	`Salesforce/SFR-Embedding-Mistral` (Meng et al., 2024)	Mistral-7B	7.1B	4096	M, C
`nemotron`	`nvidia/llama-embed-nemotron-8b` (Babakhin et al., 2025)	Llama-3.1-8B	7.5B	4096	M, C
`kalm`	`tencent/KaLM-Embedding-Gemma3-12B-2511` (Zhao et al., 2026)	Gemma-3-12B	11.8B	3840	M, S, C
`openai`	`text-embedding-3-small` (OpenAI, 2024)	API	—	1536	M, C
`mistral`	`mistral-embed` (Mistral AI, 2023)	API	—	1024	M, C

object: an embedding translator applied to only part of a document index. It measures retrieval over the resulting mixed index against a fully native target-model index and reports results separately for translated and native documents. LAReQA also ranks candidates of different provenance in a shared pool, evaluating cross-lingual answer retrieval (Roy et al., 2020). In ETBench, the candidates differ in how their embeddings were produced: translated source-model vectors compete with native target-model vectors under target-encoded queries.

Translation methods as benchmark baselines. Paired alignment methods include orthogonal Procrustes and canonical correlation analysis (Schönemann, 1966; Faruqui & Dyer, 2014). Relative representations use similarities to shared anchors (Moschella et al., 2023); Gromov–Wasserstein matches within-space geometry (Alvarez-Melis & Jaakkola, 2018). Direct translation methods include Embedding-Converter, LA2M, and HMoE (Yoon & Arik, 2025; Yang et al., 2025; Yang & Cao, 2026). HMoE already evaluates multi-model mixing. ETBench complements that work by comparing the mixed index against a fully native target index and attributing relevant results to translated and native document shards. Query-side alignment to a fixed document index (Maystre et al., 2025) and compatible representation training (Shen et al., 2020; Jaeckle et al., 2023) use different deployment contracts.

Under one split and scoring protocol, ETBench compares the identity control; unpaired `vec2vec` (Jha et al., 2025) and Gromov–Wasserstein; paired Procrustes, CCA, and Relative Encoding; a direct MLP; and Embedding-Converter, LA2M, and HMoE. Each method receives both the held-out correspondence test and the full mixed-index retrieval test (Section E).

C Encoders

Table A2 lists the 18 encoders of the benchmark, grouped by family so that the size ladders read down the table: BGE, GTE and Qwen3 each contribute a small-to-large series, and four start from the Mistral-7B-v0.1 backbone. Each is the public checkpoint named in the table, cited as its authors ask. Parameter counts are read from the released weights, so they can differ from the size in a model's name (`nemotron` is 7.5B, not 8B); dimensions are the widths of the released arrays. Backbones decide the relation a pair is filed under in Table A3.

Table A3: The 24 ordered encoder pairs of Section 4 (RQ1), with the on-disk widths the map must bridge. Groups are the stratification of Table A32.

source	target	d_s	d_t	group
bge-small	bge-large	384	1024	same family, different size
gte-small	gte-large	384	1024	same family, different size
qwen3-0.6b	qwen3-8b	1024	4096	same family, different size
qwen3-8b	qwen3-0.6b	4096	1024	same family, different size
bge-large	bge-m3	1024	1024	same family, different size
e5-large-v2	me5-large	1024	1024	same family, different size
gritlm	nv-embed-v2	4096	4096	shared backbone
linq	sfr	4096	4096	shared backbone
linq	gritlm	4096	4096	shared backbone
nv-embed-v2	gritlm	4096	4096	shared backbone
bge-large	gte-large	1024	1024	different family
bge-large	e5-large-v2	1024	1024	different family
gte-small	bge-small	384	384	different family
qwen3-8b	nemotron	4096	4096	different family
gte	nv-embed-v2	3584	4096	different family
gte	qwen3-8b	3584	4096	different family
kalm	qwen3-8b	3840	4096	different family
bge-large	nv-embed-v2	1024	4096	different family
nv-embed-v2	bge-large	4096	1024	different family
qwen3-0.6b	bge-large	1024	1024	different family
nemotron	bge-small	4096	384	different family
openai	gte	1536	3584	different family
openai	bge-large	1536	1024	different family
mistral	openai	1024	1536	different family

D DATASETS

We use the text retrieval corpora of BEIR Thakur et al. (2021). The release covers fifteen of them, from nfcorpus at 3,633 documents to msmarco at 8.8M, each with its query set and qrels. Table A4 lists the four of the main grid and the two at the million-document scale, fever and msmarco. Every corpus is stored row-aligned across encoders, so any ordered pair of encoders can be read off the same corpus without re-encoding, and the counts below are recovered from the published arrays rather than quoted from BEIR. The main grid of Section 4 (RQ1) runs on the four corpora above the rule, which are the ones whose native retrieval score is high enough for a retention ratio to be stable; on each of them D0 takes half the documents and D1 and D2 a quarter each, partitioned once per corpus and reused by every method and every encoder pair.

Table A4: BEIR retrieval corpora. Documents and queries are counted from the published arrays, not quoted from BEIR. The four above the rule form the main grid of Section 4 (RQ1).

corpus	documents	queries
nfcorpus	3,633	323
scifact	5,183	300
arguana	8,674	1,406
fiqa	57,638	648
fever	5,416,568	6,666
msmarco	8,841,823	6,980

Clustering data. The clustering experiment of Section 4 (RQ4) uses the 8 MTEB clustering tasks with gold labels in Table A5; .v2 tasks are the bootstrap-resampled fast variants.

Table A5: MTEB clustering tasks (items = embedded texts on disk). All carry gold cluster labels; several are non-English (ml).

task	items	note
BigPatentClustering.v2	2,048	long patents
CLSClusteringP2P.v2	2,048	Chinese (ml)
HALClusteringS2S.v2	2,048	French (ml)
PlscClusteringP2P.v2	2,048	Polish (ml)
RomaniBibleClustering	2,048	Romani (ml)
AlloProfClusteringS2S.v2	2,556	French (ml)
WikiCitiesClustering	3,531	city articles
SIB200ClusteringS2S	1,004	multilingual (ml)

E TRANSLATION METHODS

We use the following 10 methods in our evaluation, where a method was tuned, the choice was made on anchors held out from `D0`. The configurations below are those behind Table 1, in the order of Section 3.

Identity (`identity`). The control applies no map: each source row is placed in the target index as it is. Where the two encoders differ in width, the narrower space is zero-padded to the wider one. It has no hyperparameters, no fitting cost and no randomness beyond the common split.

`vec2vec` (Jha et al., 2025). An unsupervised translator learns space-specific adapters around a shared latent network, trained with reconstruction, cycle-consistency, vector-space-preservation and adversarial losses. Its source and target training rows come from disjoint halves of `D0`, so it never sees a paired row. Latent and hidden widths are 1024, adapter depth 3 and shared depth 4, with layer normalisation, batch size 256 and 300 epochs. The tuned variant selects the generator learning rate from $\{2 \times 10^{-5}, 6 \times 10^{-5}, 2 \times 10^{-4}, 6 \times 10^{-4}\}$ by its margin over a shuffled target on held-out anchors, and sets the discriminator rate to half of it. The unpaired split, initialisation and minibatch order are seeded with 42.

Gromov–Wasserstein (`GW`) (Peyré et al., 2016; Alvarez-Melis & Jaakkola, 2018). Entropic Gromov–Wasserstein alignment couples the two anchor clouds by comparing within-cloud distances, without using their row correspondence. The row uses 256 anchors, an absolute entropic regulariser $\epsilon = 0.001$, the squared loss, no initial pairing, and a barycentric projection that assigns each new point to its 32 nearest anchors. This is the configuration fixed before any tuning. The anchor subsample is drawn with seed 42, and the solver starts from no random coupling.

Orthogonal Procrustes (`procrustes`). The estimator is the centred orthogonal least-squares map (Schönemann, 1966). Rows are ℓ_2-normalised, and a single scalar scale is fitted jointly with the rotation. The optimal rotation is obtained in closed form from the SVD of the cross-covariance matrix, after which the scale is computed analytically. All of the computation are based on the `D0` parts of the datasets.

Canonical correlation analysis (`cca`). CCA finds maximally correlated linear coordinates in the two spaces (Faruqui & Dyer, 2014); the target is then regressed on the retained source canonical variates. The row uses the closed-form regularised solver rather than the iterative NIPALS solver, with column standardisation, 128 components and a relative ridge of 10^{-3} on both covariance matrices. The ridge only keeps the covariances invertible: 10^{-4} and 10^{-2} give the same scores to three decimals. The fit is two economy decompositions and one canonical cross-decomposition, with no random initialisation, on `openai`→`gte`.

Relative Encoding (`relative`) (Moschella et al., 2023). A source row is described by its cosine similarities to the source anchors, a softmax at temperature 0.05 is taken over its $\{64, 128, 256, 512, 1024\}$ nearest anchors by grid search, and the weights are decoded through the corresponding target anchors, with the output matched to the mean target anchor length. These settings are fixed, not searched. Fitting only stores the paired anchors, and the method has no randomness

beyond the common split.

MLP (`mlp`). A supervised baseline regresses target anchors on source anchors with a four-layer SELU perceptron and a weighted objective of mean squared error plus cosine distance, the cosine term at weight 0.5. It uses hidden width 2048, batch size 1024, AdamW at learning rate 10^{-4}, and at most 50 epochs with early stopping at patience 10 on a 0.1 fraction of `D0` held out for validation. Width and batch size were fixed before tuning. Initialisation, the validation split and minibatch order are seeded with 42.

Embedding-Converter (`embconv`) (Yoon & Arik, 2025). A supervised perceptron is trained with an absolute-error regression term and global and local cosine-distance preservation terms. It uses four SELU layers of width $\min(5d_t, 4096)$ for target width d_t, a local neighbourhood of 100, batch size 1024, Adam at learning rate 10^{-3}, at most 600 epochs, and normalised outputs. The tuned variant selects the two preservation weights, set equal, from $\{0.01, 0.03, 0.1, 0.3, 1.0\}$ by target cosine on a disjoint 0.1 fraction of `D0`, and chooses the epoch budget on the same held-out anchors. Initialisation, the split and minibatch order are seeded with 42.

LA2M (`la2m`) (Yang et al., 2025). It groups the reference pairs into small, possibly overlapping clusters of nearby vectors, sized by a maximum-likelihood estimate of the intrinsic dimension; within each cluster it projects the source and target references separately by SVD to that dimension, fits one isometry between them, and reconstructs the mapped vectors in the target space. A non-reference vector is mapped by the isometry of the cluster whose centroid is nearest to it, so the method has no global map, we search parameter of cluster size from 1, 15, 20, 40, 60, 100, 200, 500, 1000, all random seeds are set to 42.

HMoE (`hmoe`) (Yang & Cao, 2026). A hierarchical mixture of experts fits a shared base translator and routes each point through local low-rank experts. The row uses a three-level binary hierarchy, LoRA rank 8 with scale 16, a four-layer expert of width 2048, 600 base epochs, 80 LoRA epochs and 50 expert epochs, learning rate 10^{-4} and batch size 1024. For the benchmark the base loss combines mean squared error and cosine distance and the expert outputs are normalised. The split, expert initialisation and minibatch order are seeded with 42.

F KV-CACHE SETUP

Prior work reuses cached context within a model by combining cached chunks with selective recomputation (Yao et al., 2025). Across models, Heo et al. (2026) transfer the cache within an LLM family with closed-form ridge maps fitted per head, after removing RoPE from the keys. We follow that construction and compare the translation methods of Table 2 against the target model's own cache.

The source is `Qwen3-4B` and the target is the generative `Qwen3-8B`. Both have 36 layers, 8 KV heads per layer, head dimension 128 and `rope_theta` $= 10^6$, so each source head maps directly onto the target head at the same layer and index. Keys are extracted with RoPE removed, and the target rotation is re-applied at each prefix position on installation; the two rotations are exact inverses up to the precision in which keys are stored. Values carry no positional term and are left unchanged. Each method fits one map per target $(\text{layer}, \text{head}, K/V)$ block, $36 \times 8 \times 2 = 576$ blocks in total, reading only the corresponding source block; `identity` fits nothing.

Maps are fitted on the first 160 qualifying documents of the `sample-10BT` training split of FineWeb-Edu, each truncated to $1{,}024$ tokens; shorter documents are skipped and the order is not shuffled. The next 40 qualifying documents are held out, and each is evaluated at prefix lengths 128, 256, 384 and 512, giving 160 document–prefix observations at every scored position.

For each prefix, we install the translated and the native cache in separate passes and feed the target the same true continuation tokens in both. Depth-1 next-token agreement is the share of document–prefix pairs whose two logit vectors at the first scored position have the same argmax; since the first continuation token is fed before this logit is produced, the logit predicts the token after it. Depth-16 agreement averages the same indicator over the first 16 scored positions. KL is $D_{\mathrm{KL}}(p_{\text{native}} \,\|\, p_{\text{translated}})$ between the two softmax distributions, averaged over the scored positions of each prefix and then equally over the four prefix lengths.

Table A6: Cost per method. Each map is fitted on 65,536 `fever` anchor pairs (`gte→kalm`) and then applied to 1,048,576 `D1` rows; fit and transform are each the mean of five runs (± one standard deviation). Times are wall clock and exclude data loading, which is the same for every method: about 3 s for the anchors and 25 s for the 1M rows. CPU rows use eight threads; GPU rows one H100 with eight threads.

		Time (s)		Peak memory (GB)	
method	device	fit, 64K	transform, 1M	RAM	VRAM
`identity`	CPU	0.0 ± 0.0	0.0 ± 0.0	15	–
`relative`	CPU	1.0 ± 0.0	3423.4 ± 310.5	63	–
`procrustes`	CPU	7.6 ± 0.1	21.3 ± 3.5	46	–
`cca`	CPU	137.4 ± 3.9	147.5 ± 37.2	106	–
`GW`	CPU	11.5 ± 0.2	2358.0 ± 289.9	46	–
`mlp`	H100	46.1 ± 82.8	24.0 ± 0.5	61	3
`vec2vec`	H100	986.3 ± 69.9	12.5 ± 0.8	31	2
`embconv`	H100	116.7 ± 0.9	17.0 ± 1.7	31	4
`la2m`	H100	259.1 ± 5.4	84.3 ± 4.1	64	14
`hmoe`	H100	682.5 ± 3.6	51.7 ± 2.5	68	69

G Efficiency Comparison

Table A6 reports the cost of each configuration in Appendix E at a scale that makes the differences apparent: 65,536 anchor pairs from `fever` for fitting and one million `D1` rows for transformation. The closed-form methods fit quickly: `relative`, `procrustes`, and `GW` finish within 12 s, while `cca` takes 137 s. Training-based methods take from 46 s for `mlp`, whose fit time varies widely across runs, to 986 s for `vec2vec`. Fitting `hmoe` is also time-consuming, 683 s, because it requires training multiple LoRA adapters.

Transformation exposes a larger difference. Methods that apply a fitted map, global or local, process one million rows in 12–148 s; `la2m` and `hmoe`, which route each row to a local map, take 84 s and 52 s. Methods that consult anchors at inference incur a per-row cost: `GW` and `relative` take 39 and 57 minutes, respectively, on the CPU. Memory use also varies substantially: the float64 decompositions in `cca` require 106 GB of host memory, and `hmoe` is the only method that occupies most of an 80 GB GPU (69 GB); the next largest, `la2m`, uses 14 GB.

H Mixed-population clustering

We assess whether translated vectors integrate into clusters containing native vectors when both coexist, as in a staged migration. The evaluation covers six encoder pairs (three shared-backbone and three cross-family) and eight labelled MTEB tasks. For each task, `D0` (half the items) provides anchor pairs for fitting the map; translated `D1` and native `D2` each contain a quarter. We run k-means on the unit-normalised mixture of `D1` and `D2`, with k set to the number of gold classes. Table A7 reports averages over the 48 pair–task cells. Hits@1 and cosine measure alignment between translated `D1` vectors and their native counterparts. Centroid agreement measures how often the two versions of an item are assigned to the same centroid fitted on native `D2`. Herding is Cramér's V between mixture cluster assignments and translation status, minus its mean under shuffled status labels. To measure retention of class structure, we also cluster native `D1` together with the same `D2`. Let C_{mix} and C_{native} denote the cluster assignments from these two runs, and let Y denote the gold labels of the items in `D1` and `D2`. AMI retention is $\text{AMI}(C_{\text{mix}}, Y)/\text{AMI}(C_{\text{native}}, Y)$, where

$$\text{AMI}(C, Y) = \frac{I(C;Y) - \mathbb{E}_0[I(C;Y)]}{\frac{1}{2}\{H(C) + H(Y)\} - \mathbb{E}_0[I(C;Y)]}.$$

Here I is mutual information, H is entropy, and $\mathbb{E}_0$ is the expectation over random assignments with the observed cluster and class sizes held fixed. The details are shown in the Tab. A7.

I The per-pair results of Table 1

Table 1 averages 24 ordered pairs over four BEIR corpora. The tables below show every cell, with one table per pair, one block per corpus and all nine retrieval columns; `procrustes` is the fit

Table A7: Mixed-population clustering over six encoder pairs and eight MTEB tasks. Vector Hits@1 retrieves each translated record's native counterpart; cos compares the matched vectors. Centroid agreement asks whether the translation reaches the native record's centroid. Herding measures how much a record's cluster depends on whether it is a translated `D1` record or a native `D2` record. AMI retention is relative to native target clustering. Higher is better except for herding.

	vector alignment		clustering outcome		
mapper	Hits@1	cos	centroid agreement	herding ↓	AMI ret.
`identity`	.499	.276	.345	.867	.754
`procrustes`	.990	.783	.647	.204	.964
`cca`	.991	.735	.204	.755	.445
`GW`	.983	**.791**	.238	.649	.593
`vec2vec`	.002	.741	.108	.864	.516
`embconv`	**.992**	.763	.411	.684	.600
`la2m`	.968	.771	**.783**	.056	**.999**
`hmoe`	.917	.718	.776	**.021**	.966

Centroids are fitted on `D2`. The centroid-agreement floor is the majority share, .140; zero is the herding floor.

used in Table 1. Every entry is a *retention* ratio. As in Table 1, `D0` contains half the corpus and the anchors used to fit the map; `D1` and `D2` contain a quarter each. For each pair, corpus and method, the mixed index uses target-model vectors except on `D1`, where source-model vectors pass through the fitted map. Queries use the target model. Retention divides the mixed-index score by the score of a target-only index over the same documents and queries. Thus 1 is parity; corpus difficulty cancels because only the origin of `D1`'s vectors changes. The three columns under each metric score those same runs against different answer sets: *full* uses the whole index, `D1` the translated documents and `D2` the native documents, with numerator and denominator evaluated on the same slice. On `D2`, the ideal is 1: a higher ratio means translated rows failed to compete for ranks. The results are shown from Tables A8 to A31.

Table A8: **bge-small→bge-large**, 384 → 1024. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.526	.000	1.045	.539	.000	1.047	.529	.000	1.020
relative	.782	.817	.766	.765	.756	.785	.879	.966	.792
procrustes	.631	.442	.810	.617	.360	.848	.866	.925	.806
cca	.867	.870	.860	.846	.809	.872	.979	.986	.973
GW	.616	.368	.845	.605	.294	.884	.815	.888	.738
mlp	.860	.905	.804	.850	.857	.829	.888	.993	.779
vec2vec	.470	.003	.925	.496	.001	.955	.448	.078	.779
embconv	.857	.703	1.012	.821	.627	1.018	.970	.949	.993
la2m	.916	.804	1.025	.889	.745	1.030	.993	.993	.993
hmoe	.939	.914	.964	.925	.872	.979	.993	1.000	.986
NFCorpus									
identity	.826	.000	1.172	.902	.000	1.150	.738	.000	1.087
relative	.796	.893	.789	.898	.828	.860	.777	1.376	.519
procrustes	.681	.448	.818	.827	.433	.878	.693	1.067	.514
cca	.876	.856	.879	.938	.795	.914	.896	1.267	.720
GW	.743	.352	.917	.874	.299	.983	.665	1.069	.514
mlp	.853	.946	.799	.932	.839	.865	.777	1.407	.489
vec2vec	.760	.017	1.062	.867	.016	1.075	.449	.247	.528
embconv	.901	.624	1.030	.932	.628	1.036	.968	1.044	.956
la2m	.937	.668	1.074	.953	.629	1.080	.982	.985	.998
hmoe	.915	.799	.961	.940	.744	.987	.974	1.152	.887
ArguAna									
identity	.583	.000	1.168	.639	.000	1.280	.502	.000	1.004
relative	.781	1.119	.442	.773	1.144	.400	.942	1.004	.879
procrustes	.648	.460	.837	.650	.379	.922	.909	.963	.856
cca	.916	1.013	.819	.899	.995	.802	.987	1.006	.968
GW	.610	.274	.948	.633	.210	1.057	.863	.899	.826
mlp	.864	.999	.730	.856	.963	.748	.938	1.006	.871
vec2vec	.582	.000	1.167	.638	.000	1.279	.504	.004	1.004
embconv	.954	.960	.948	.943	.943	.943	.999	1.001	.997
la2m	.963	.926	1.000	.954	.902	1.006	1.000	1.001	.999
hmoe	.967	1.023	.911	.955	1.013	.897	1.000	1.006	.994
FiQA									
identity	.647	.000	1.101	.716	.000	1.105	.611	.000	1.018
relative	.699	.735	.682	.733	.644	.727	.781	1.008	.626
procrustes	.582	.174	.850	.643	.128	.894	.673	.636	.683
cca	.807	.803	.793	.837	.758	.825	.902	1.050	.805
GW	.623	.032	1.036	.701	.027	1.058	.675	.463	.806
mlp	.835	.752	.880	.851	.672	.920	.887	1.099	.757
vec2vec	.647	.000	1.101	.716	.000	1.105	.611	.000	1.018
embconv	.894	.867	.914	.906	.836	.929	.978	1.002	.967
la2m	.888	.679	1.038	.887	.624	1.051	.967	.911	1.002
hmoe	.884	.952	.845	.896	.926	.859	.952	1.044	.890

Table A9: **gte-small→gte-large**, 384 → 1024. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.521	.000	1.044	.533	.000	1.056	.532	.000	1.020
relative	.740	.942	.559	.749	.925	.566	.805	.973	.640
procrustes	.621	.563	.677	.595	.485	.698	.805	.953	.660
cca	.894	.990	.798	.897	.980	.808	.925	.993	.863
GW	.629	.557	.702	.617	.490	.738	.771	.915	.620
mlp	.842	1.008	.675	.854	1.001	.696	.819	1.000	.640
vec2vec	.456	.003	.903	.479	.001	.938	.408	.054	.730
embconv	.859	.726	1.001	.822	.645	1.010	.974	.963	.987
la2m	.951	.867	1.032	.931	.821	1.040	.997	.993	1.000
hmoe	.945	1.014	.879	.944	1.005	.880	.976	1.000	.953
NFCorpus									
identity	.820	.000	1.177	.898	.000	1.190	.735	.000	1.097
relative	.631	1.163	.374	.765	1.197	.404	.608	1.327	.297
procrustes	.496	.537	.468	.634	.535	.543	.552	1.063	.311
cca	.813	1.135	.670	.907	1.132	.726	.823	1.322	.560
GW	.547	.699	.473	.708	.681	.562	.500	1.091	.263
mlp	.723	1.353	.407	.829	1.275	.439	.605	1.372	.240
vec2vec	.501	.050	.735	.669	.055	.842	.301	.299	.307
embconv	.864	.745	.931	.913	.751	.965	.943	1.074	.882
la2m	.946	.706	1.060	.949	.632	1.095	.984	.990	.979
hmoe	.892	1.138	.799	.931	1.112	.822	.921	1.324	.705
ArguAna									
identity	.592	.000	1.181	.642	.000	1.286	.501	.000	1.004
relative	.778	1.150	.408	.784	1.177	.391	.916	1.003	.829
procrustes	.653	.438	.866	.661	.364	.960	.892	.953	.832
cca	.925	1.017	.833	.921	1.001	.842	.987	1.001	.973
GW	.669	.407	.930	.684	.329	1.040	.879	.957	.800
mlp	.859	.914	.805	.848	.837	.859	.917	1.003	.832
vec2vec	.592	.000	1.181	.642	.000	1.286	.501	.000	1.004
embconv	.959	.981	.937	.954	.972	.936	1.001	1.001	1.000
la2m	.976	.938	1.013	.971	.917	1.025	1.000	1.000	1.000
hmoe	.964	1.042	.886	.958	1.039	.877	.996	1.001	.991
FiQA									
identity	.652	.000	1.100	.723	.000	1.103	.625	.000	1.037
relative	.651	.932	.475	.703	.895	.505	.687	1.023	.464
procrustes	.567	.296	.758	.629	.246	.807	.637	.753	.551
cca	.832	.930	.769	.881	.913	.800	.860	1.036	.731
GW	.604	.226	.869	.669	.180	.922	.660	.727	.592
mlp	.816	.872	.777	.855	.803	.822	.794	1.076	.598
vec2vec	.650	.000	1.096	.722	.000	1.101	.579	.004	.962
embconv	.908	.868	.924	.924	.863	.943	.976	.994	.957
la2m	.894	.678	1.040	.902	.642	1.053	.965	.921	.992
hmoe	.900	1.048	.798	.930	1.050	.812	.938	1.059	.856

Table A10: **qwen3-0.6b→qwen3-8b**, 1024 → 4096. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.522	.000	1.035	.533	.000	1.043	.540	.000	1.020
relative	.886	.887	.906	.871	.840	.916	.971	.990	.954
procrustes	.675	.380	.966	.642	.281	.984	.952	.980	.928
cca	.887	.798	.977	.863	.737	.983	.993	1.007	.980
GW	.635	.277	.982	.618	.206	1.002	.933	.952	.922
mlp	.913	.850	.972	.896	.800	.981	.986	1.007	.967
vec2vec	.522	.000	1.035	.533	.000	1.043	.540	.000	1.020
embconv	.771	.511	1.027	.725	.409	1.035	.974	.956	.994
la2m	.925	.827	1.020	.907	.782	1.028	1.000	1.007	.994
hmoe	.931	.857	1.003	.913	.811	1.013	1.000	1.007	.994
NFCorpus									
identity	.820	.000	1.156	.912	.000	1.171	.740	.000	1.091
relative	.904	.942	.923	.940	.890	.954	.928	1.256	.787
procrustes	.808	.308	1.010	.902	.262	1.024	.810	1.046	.696
cca	.922	.737	1.011	.955	.632	1.029	.981	1.193	.884
GW	.831	.205	1.086	.910	.156	1.122	.799	1.070	.682
mlp	.920	.753	1.002	.963	.671	1.025	.955	1.198	.843
vec2vec	.820	.000	1.156	.912	.000	1.171	.738	.002	1.086
embconv	.876	.413	1.079	.938	.370	1.109	.958	.915	.988
la2m	.959	.652	1.095	.954	.536	1.110	.986	.940	1.011
hmoe	.943	.705	1.055	.945	.594	1.079	1.008	1.059	.979
ArguAna									
identity	.567	.000	1.149	.616	.000	1.253	.500	.000	1.000
relative	.937	1.056	.814	.945	1.107	.778	.998	.997	.999
procrustes	.660	.230	1.102	.676	.165	1.204	.976	.959	.994
cca	.921	.792	1.054	.911	.708	1.120	.999	1.000	.997
GW	.659	.205	1.126	.673	.133	1.232	.984	.979	.989
mlp	.954	.906	1.003	.944	.860	1.030	.998	1.000	.996
vec2vec	.567	.000	1.149	.616	.000	1.253	.500	.000	1.000
embconv	.940	.807	1.077	.942	.745	1.146	.998	.997	.999
la2m	.961	.903	1.020	.949	.869	1.032	.999	1.000	.999
hmoe	.965	.927	1.005	.952	.891	1.015	.999	1.000	.999
FiQA									
identity	.658	.000	1.091	.750	.000	1.101	.612	.000	1.007
relative	.769	.813	.762	.804	.774	.776	.916	.987	.857
procrustes	.667	.079	1.047	.755	.062	1.073	.827	.673	.913
cca	.838	.603	.979	.855	.511	1.010	.947	.966	.917
GW	.655	.001	1.086	.749	.001	1.100	.584	.015	.949
mlp	.897	.768	.968	.912	.707	.989	.968	1.012	.927
vec2vec	.658	.000	1.091	.750	.000	1.101	.612	.000	1.007
embconv	.887	.698	1.002	.909	.655	1.013	.981	.967	.985
la2m	.892	.614	1.059	.915	.564	1.072	.975	.919	.998
hmoe	.912	.816	.961	.927	.770	.975	.976	.986	.958

Table A11: **qwen3-8b→qwen3-0.6b**, 4096 → 1024. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.515	.000	1.034	.532	.000	1.040	.530	.000	1.020
relative	.892	.932	.882	.882	.907	.883	.930	.980	.891
procrustes	.658	.405	.911	.633	.314	.925	.895	.925	.878
cca	.940	.907	.981	.922	.866	.984	.976	.986	.966
GW	.667	.370	.951	.646	.286	.973	.885	.935	.844
mlp	.985	1.019	.951	.981	.999	.957	.956	1.000	.918
vec2vec	.515	.000	1.034	.532	.000	1.040	.530	.000	1.020
embconv	.827	.663	1.008	.787	.573	1.018	.960	.925	.993
la2m	.923	.826	1.017	.903	.781	1.023	.991	.983	1.000
hmoe	.974	.949	1.004	.958	.914	1.007	.986	.980	.993
NFCorpus									
identity	.816	.000	1.172	.904	.000	1.205	.737	.000	1.112
relative	.891	.985	.914	.919	.994	.946	.972	1.323	.838
procrustes	.783	.378	.992	.885	.355	1.026	.852	1.142	.719
cca	.954	.906	.995	.964	.841	1.009	1.016	1.226	.940
GW	.799	.238	1.047	.886	.205	1.087	.809	1.026	.718
mlp	.949	.964	.955	.970	.871	.978	1.004	1.364	.857
vec2vec	.816	.000	1.172	.904	.000	1.205	.737	.006	1.106
embconv	.862	.490	1.057	.916	.493	1.091	.988	1.042	.970
la2m	.952	.652	1.103	.952	.556	1.138	1.007	.968	1.027
hmoe	.973	.849	1.044	.969	.782	1.051	1.018	1.098	.982
ArguAna									
identity	.572	.000	1.171	.627	.000	1.283	.500	.000	1.001
relative	.971	1.129	.806	.999	1.211	.776	.994	.997	.991
procrustes	.671	.251	1.111	.697	.195	1.223	.961	.943	.980
cca	.955	.839	1.077	.957	.768	1.154	.999	.999	.999
GW	.664	.205	1.145	.687	.142	1.259	.974	.974	.973
mlp	1.005	1.048	.960	1.019	1.063	.972	.996	1.000	.991
vec2vec	.572	.000	1.171	.627	.000	1.283	.500	.000	1.001
embconv	1.005	1.010	1.000	1.022	1.021	1.023	1.000	1.000	1.000
la2m	1.002	.961	1.045	1.002	.934	1.073	1.001	1.000	1.001
hmoe	1.019	1.049	.989	1.032	1.073	.990	1.000	1.000	1.000
FiQA									
identity	.649	.000	1.104	.730	.000	1.109	.615	.000	1.040
relative	.837	1.140	.708	.851	1.091	.739	.923	1.115	.806
procrustes	.686	.201	1.033	.748	.147	1.061	.810	.778	.836
cca	.949	1.031	.922	.953	.972	.953	.971	1.099	.891
GW	.639	.001	1.085	.726	.002	1.098	.558	.015	.924
mlp	1.029	1.234	.917	1.036	1.211	.937	.990	1.122	.916
vec2vec	.649	.000	1.104	.730	.000	1.109	.615	.000	1.040
embconv	.995	1.032	.987	.984	1.012	.994	1.005	1.046	.985
la2m	.974	.892	1.039	.961	.850	1.047	1.014	1.026	1.011
hmoe	1.030	1.187	.946	1.031	1.179	.959	1.006	1.085	.965

Table A12: **bge-large→bge-m3**, 1024 → 1024. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.529	.000	1.040	.545	.000	1.049	.520	.000	1.018
relative	.702	.816	.620	.700	.779	.628	.826	.937	.722
procrustes	.423	.199	.647	.410	.142	.669	.740	.764	.729
cca	.821	.803	.840	.795	.766	.829	.931	.961	.897
GW	.371	.005	.717	.394	.004	.741	.461	.270	.648
mlp	.799	.944	.660	.802	.920	.677	.868	1.014	.729
vec2vec	.529	.000	1.040	.545	.000	1.049	.516	.000	1.011
embconv	.771	.548	.995	.724	.447	.996	.926	.884	.971
la2m	.884	.729	1.030	.853	.657	1.038	.971	.951	.993
hmoe	.916	.874	.963	.896	.846	.949	.974	.986	.963
NFCorpus									
identity	.825	.000	1.147	.915	.000	1.163	.742	.000	1.067
relative	.714	1.018	.600	.807	1.075	.569	.740	1.422	.481
procrustes	.558	.382	.651	.693	.383	.657	.698	1.158	.479
cca	.814	.715	.892	.890	.702	.907	.896	1.338	.678
GW	.606	.146	.849	.763	.145	.875	.513	.727	.491
mlp	.768	1.104	.610	.824	.994	.636	.786	1.563	.467
vec2vec	.825	.000	1.147	.915	.000	1.163	.743	.004	1.067
embconv	.856	.451	1.033	.915	.474	1.058	.943	.952	.932
la2m	.937	.574	1.091	.960	.522	1.120	1.012	1.051	.974
hmoe	.921	.718	1.023	.952	.683	1.045	.982	1.146	.888
ArguAna									
identity	.593	.000	1.195	.643	.000	1.299	.498	.000	1.006
relative	.878	1.305	.445	.917	1.409	.415	.930	1.004	.855
procrustes	.626	.390	.866	.648	.333	.971	.862	.894	.830
cca	.995	1.131	.857	1.006	1.140	.871	.986	.999	.972
GW	.498	.000	1.004	.554	.000	1.120	.409	.026	.799
mlp	.942	1.161	.720	.963	1.178	.744	.931	1.003	.858
vec2vec	.561	.001	1.128	.617	.001	1.245	.449	.036	.870
embconv	1.012	1.087	.937	1.027	1.109	.944	1.000	1.003	.997
la2m	.979	.902	1.057	.971	.856	1.087	.999	.997	1.000
hmoe	1.025	1.101	.948	1.036	1.120	.950	.999	1.003	.996
FiQA									
identity	.649	.000	1.085	.702	.000	1.095	.630	.000	1.038
relative	.718	.867	.663	.750	.775	.709	.803	1.082	.632
procrustes	.548	.094	.865	.604	.064	.906	.606	.501	.673
cca	.820	.844	.818	.841	.795	.844	.921	1.095	.810
GW	.627	.008	1.045	.688	.005	1.072	.562	.142	.837
mlp	.881	.926	.876	.901	.851	.922	.924	1.166	.761
vec2vec	.649	.000	1.085	.702	.000	1.095	.630	.000	1.038
embconv	.925	.897	.949	.921	.854	.959	.999	1.034	.974
la2m	.879	.620	1.054	.879	.573	1.067	.983	.941	1.007
hmoe	.926	1.001	.895	.937	.971	.912	.985	1.093	.902

Table A13: **e5-large-v2→me5-large**, 1024 → 1024. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.530	.000	1.038	.550	.000	1.040	.530	.000	1.021
relative	.545	.798	.324	.533	.751	.323	.689	.972	.416
procrustes	.271	.188	.353	.257	.147	.359	.633	.819	.444
cca	.759	.872	.641	.742	.852	.623	.857	.972	.755
GW	.289	.175	.394	.275	.133	.403	.566	.767	.374
mlp	.653	.983	.347	.658	.967	.352	.691	1.007	.381
vec2vec	.290	.006	.547	.308	.002	.572	.283	.119	.416
embconv	.827	.678	.968	.798	.595	.975	.959	.948	.972
la2m	.874	.716	1.015	.844	.643	1.021	.996	.993	1.000
hmoe	.887	1.008	.768	.877	.996	.756	.923	1.014	.839
NFCorpus									
identity	.815	.000	1.175	.908	.000	1.252	.740	.000	1.106
relative	.330	.932	.059	.440	.924	.053	.396	1.256	.035
procrustes	.166	.384	.071	.255	.480	.071	.325	.972	.035
cca	.457	.939	.210	.580	.997	.212	.566	1.274	.248
GW	.138	.282	.071	.194	.308	.071	.280	.892	.029
mlp	.360	1.032	.057	.512	1.074	.050	.488	1.430	.032
vec2vec	.099	.082	.089	.163	.102	.098	.138	.390	.035
embconv	.657	.679	.664	.731	.679	.686	.857	1.067	.771
la2m	.927	.592	1.090	.935	.514	1.157	.984	.936	1.000
hmoe	.627	1.197	.353	.743	1.199	.345	.777	1.429	.461
ArguAna									
identity	.590	.000	1.200	.634	.000	1.299	.501	.000	1.006
relative	.688	1.238	.118	.718	1.298	.109	.679	1.012	.344
procrustes	.277	.245	.310	.284	.218	.352	.539	.698	.379
cca	.787	1.120	.442	.790	1.142	.422	.893	1.007	.779
GW	.396	.427	.363	.405	.387	.424	.593	.870	.314
mlp	.733	1.222	.227	.760	1.259	.236	.673	1.012	.332
vec2vec	.440	.003	.893	.491	.002	1.003	.370	.038	.705
embconv	.944	.987	.899	.942	.979	.904	.996	1.004	.987
la2m	.942	.845	1.042	.927	.798	1.061	.999	.999	1.000
hmoe	.915	1.161	.660	.913	1.199	.613	.960	1.010	.910
FiQA									
identity	.632	.000	1.088	.699	.000	1.088	.606	.000	1.049
relative	.382	.731	.162	.434	.703	.183	.466	.965	.121
procrustes	.147	.081	.194	.161	.065	.218	.250	.392	.121
cca	.503	.884	.243	.575	.894	.268	.560	.992	.248
GW	.194	.007	.327	.235	.007	.372	.108	.013	.161
mlp	.531	.998	.210	.614	1.026	.232	.517	1.034	.136
vec2vec	.274	.000	.455	.342	.000	.503	.143	.003	.222
embconv	.908	.946	.903	.926	.935	.912	.972	1.029	.934
la2m	.878	.676	1.030	.887	.631	1.032	.966	.923	.996
hmoe	.742	1.210	.424	.808	1.216	.435	.759	1.129	.490

Table A14: **gritlm→nv-embed-v2**, 4096 → 4096. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.791	.537	1.037	.763	.465	1.048	.987	.966	1.006
relative	.890	.852	.950	.866	.803	.945	.975	.986	.968
procrustes	.618	.218	1.018	.593	.143	1.038	.914	.884	.949
cca	.906	.805	1.008	.877	.741	1.022	.995	1.007	.987
GW	.640	.251	1.021	.610	.171	1.037	.935	.918	.949
mlp	.939	.867	1.011	.921	.819	1.026	1.003	1.007	1.000
vec2vec	.520	.000	1.039	.532	.000	1.051	.550	.024	1.006
embconv	.726	.437	1.024	.676	.328	1.039	.968	.935	1.000
la2m	.919	.806	1.028	.898	.759	1.036	1.000	1.000	1.000
hmoe	.945	.863	1.026	.923	.815	1.036	1.000	1.000	1.000
NFCorpus									
identity	.884	.256	1.154	.908	.241	1.140	.940	.667	1.056
relative	.916	.956	.937	.930	.962	.939	.956	1.067	.930
procrustes	.783	.273	1.011	.878	.266	1.016	.818	.876	.774
cca	.895	.770	.952	.911	.742	.966	.963	1.109	.939
GW	.817	.114	1.111	.872	.102	1.114	.825	.802	.857
mlp	.916	.746	.979	.913	.678	.999	.965	1.088	.938
vec2vec	.814	.000	1.172	.871	.000	1.155	.733	.000	1.075
embconv	.849	.459	1.054	.891	.449	1.066	.917	.846	.961
la2m	.913	.391	1.147	.912	.357	1.130	.963	.819	1.035
hmoe	.928	.593	1.077	.912	.539	1.076	.974	.938	1.001
ArguAna									
identity	.643	.173	1.111	.639	.110	1.173	.991	.981	1.000
relative	.929	1.157	.702	.939	1.221	.654	.996	1.001	.991
procrustes	.605	.082	1.126	.638	.057	1.226	.924	.861	.987
cca	.901	.763	1.039	.889	.664	1.115	.997	1.001	.993
GW	.576	.000	1.151	.622	.000	1.250	.510	.027	.991
mlp	.967	.955	.978	.957	.912	1.003	.997	1.001	.993
vec2vec	.578	.000	1.154	.624	.000	1.253	.505	.009	1.000
embconv	.882	.715	1.049	.870	.626	1.115	1.001	1.001	1.000
la2m	.900	.764	1.034	.860	.666	1.055	1.001	1.001	1.000
hmoe	.968	.925	1.010	.953	.882	1.024	1.001	1.001	1.000
FiQA									
identity	.664	.035	1.089	.746	.028	1.103	.705	.189	1.034
relative	.790	1.066	.650	.824	1.046	.654	.897	1.063	.786
procrustes	.577	.043	.945	.675	.032	.986	.667	.441	.804
cca	.799	.697	.883	.814	.618	.916	.931	.990	.895
GW	.627	.000	1.047	.726	.000	1.079	.551	.043	.880
mlp	.954	1.118	.871	.982	1.129	.876	.996	1.071	.939
vec2vec	.647	.000	1.090	.736	.000	1.104	.623	.000	1.034
embconv	.934	.855	1.010	.938	.826	1.025	.986	.985	.996
la2m	.919	.713	1.054	.919	.661	1.067	.977	.912	1.017
hmoe	.970	.991	.970	.983	.984	.979	.994	1.017	.979

Table A15: **linq→sfr**, 4096 → 4096. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.525	.000	1.049	.536	.000	1.058	.538	.014	1.013
relative	.742	.932	.581	.751	.915	.595	.808	.986	.637
procrustes	.558	.440	.682	.547	.370	.718	.763	.905	.630
cca	.905	1.064	.760	.915	1.068	.772	.938	1.000	.880
GW	.375	.031	.721	.396	.022	.759	.476	.318	.613
mlp	.893	1.107	.693	.908	1.117	.707	.899	1.007	.800
vec2vec	.512	.000	1.024	.529	.000	1.044	.509	.041	.933
embconv	.782	.549	1.024	.741	.457	1.032	.974	.949	1.000
la2m	.987	.996	.985	.988	.999	.979	1.000	1.007	.993
hmoe	.988	1.072	.912	.997	1.086	.914	.986	1.007	.967
NFCorpus									
identity	.824	.000	1.148	.908	.000	1.148	.733	.000	1.069
relative	.710	1.127	.531	.811	1.069	.562	.640	1.321	.326
procrustes	.574	.500	.566	.725	.451	.613	.573	1.057	.346
cca	.868	1.450	.608	.908	1.362	.595	.846	1.433	.601
GW	.469	.064	.622	.638	.074	.665	.321	.392	.277
mlp	.815	1.469	.506	.866	1.299	.508	.719	1.482	.397
vec2vec	.607	.056	.797	.766	.057	.882	.325	.297	.315
embconv	.897	.537	1.042	.917	.435	1.053	.945	.978	.944
la2m	.971	.798	1.027	.961	.701	1.037	.989	1.057	.955
hmoe	.956	1.251	.819	.960	1.189	.794	.966	1.310	.819
ArguAna									
identity	.583	.000	1.175	.625	.000	1.266	.501	.001	1.004
relative	.797	1.174	.414	.795	1.203	.376	.928	1.003	.853
procrustes	.621	.343	.902	.634	.291	.986	.877	.918	.837
cca	.877	.980	.772	.868	.938	.796	.963	1.000	.925
GW	.507	.026	.996	.552	.022	1.095	.555	.299	.814
mlp	.889	1.107	.667	.871	1.083	.653	.939	1.004	.873
vec2vec	.583	.000	1.175	.625	.000	1.266	.504	.009	1.001
embconv	.948	.912	.985	.954	.891	1.020	.998	.996	1.000
la2m	1.002	1.040	.964	.992	1.029	.955	1.000	1.001	.999
hmoe	.994	1.096	.890	.982	1.091	.871	.999	1.003	.994
FiQA									
identity	.656	.000	1.106	.736	.000	1.108	.621	.000	1.038
relative	.610	.780	.522	.665	.730	.583	.735	1.000	.552
procrustes	.570	.115	.884	.667	.107	.944	.606	.469	.694
cca	.772	.865	.718	.822	.824	.767	.822	1.036	.665
GW	.616	.000	1.042	.713	.000	1.073	.475	.007	.802
mlp	.873	1.140	.697	.920	1.106	.742	.860	1.102	.699
vec2vec	.656	.000	1.106	.736	.000	1.108	.621	.000	1.038
embconv	.951	.873	.998	.952	.836	1.006	1.005	1.036	.992
la2m	.910	.689	1.039	.924	.655	1.055	.977	.946	.999
hmoe	.971	1.120	.858	.980	1.107	.873	.998	1.087	.955

Table A16: **linq→gritlm**, 4096 → 4096. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.771	.509	1.021	.748	.455	1.025	.968	.926	1.006
relative	.869	.939	.820	.866	.925	.823	.947	.973	.928
procrustes	.686	.462	.901	.655	.379	.923	.906	.946	.876
cca	.910	.921	.903	.900	.899	.909	.971	.993	.954
GW	.509	.017	.991	.524	.012	1.008	.584	.225	.902
mlp	.617	.199	1.024	.592	.134	1.030	.900	.780	1.013
vec2vec	.517	.000	1.025	.528	.000	1.030	.544	.034	1.006
embconv	.766	.530	1.007	.725	.438	1.012	.956	.919	.994
la2m	.947	.891	.998	.932	.865	.997	.997	.993	1.000
hmoe	.922	.835	1.002	.899	.791	1.004	.992	.990	.994
NFCorpus									
identity	.928	.456	1.142	.958	.399	1.159	.928	.616	1.036
relative	.840	1.278	.722	.910	1.289	.719	.823	1.257	.655
procrustes	.687	.469	.791	.798	.398	.855	.735	1.138	.552
cca	.902	1.158	.824	.932	1.087	.852	.942	1.363	.759
GW	.767	.023	1.108	.890	.021	1.169	.484	.266	.605
mlp	.830	.072	1.171	.909	.054	1.197	.839	.332	1.042
vec2vec	.816	.000	1.177	.909	.000	1.199	.724	.000	1.048
embconv	.855	.421	1.083	.914	.419	1.105	.939	.862	.982
la2m	.986	.916	1.047	.986	.891	1.058	1.000	1.014	.997
hmoe	.955	.700	1.079	.962	.632	1.105	.990	.995	.987
ArguAna									
identity	.620	.084	1.162	.643	.053	1.253	.907	.813	1.001
relative	.877	1.284	.467	.919	1.416	.405	.986	1.000	.973
procrustes	.581	.228	.938	.607	.179	1.048	.883	.864	.903
cca	.920	1.032	.806	.930	1.004	.853	.987	1.001	.973
GW	.566	.000	1.139	.613	.000	1.247	.496	.061	.930
mlp	.626	.084	1.174	.657	.057	1.276	.865	.728	1.001
vec2vec	.585	.000	1.176	.628	.000	1.277	.522	.042	1.001
embconv	.933	.985	.880	.947	.988	.905	.998	.996	1.000
la2m	1.019	1.078	.960	1.022	1.092	.949	1.000	1.000	1.000
hmoe	1.006	1.096	.914	1.017	1.117	.914	.999	1.000	.997
FiQA									
identity	.724	.164	1.084	.792	.140	1.101	.823	.486	1.030
relative	.740	.926	.655	.771	.878	.653	.878	1.012	.780
procrustes	.634	.082	.998	.731	.057	1.035	.732	.552	.839
cca	.782	.664	.857	.806	.581	.879	.896	.986	.838
GW	.652	.000	1.087	.751	.000	1.105	.609	.003	1.010
mlp	.727	.180	1.084	.790	.147	1.099	.841	.546	1.028
vec2vec	.652	.000	1.088	.751	.000	1.105	.620	.000	1.030
embconv	.858	.670	.982	.866	.601	.987	.969	.953	.974
la2m	.935	.805	1.006	.941	.769	1.014	.980	.953	1.002
hmoe	.923	.934	.904	.924	.888	.900	.973	1.001	.958

Table A17: **nv-embed-v2→gritlm**, 4096 → 4096. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.836	.637	1.019	.818	.594	1.023	.980	.939	1.013
relative	.867	.906	.847	.859	.893	.846	.966	.983	.954
procrustes	.694	.488	.892	.672	.416	.914	.910	.953	.876
cca	.952	1.013	.898	.950	1.008	.901	.978	1.000	.961
GW	.727	.518	.926	.693	.419	.953	.930	.993	.876
mlp	.977	1.053	.907	.983	1.061	.911	.974	1.000	.954
vec2vec	.517	.000	1.025	.528	.000	1.030	.534	.000	1.013
embconv	.757	.506	1.009	.708	.392	1.016	.973	.943	1.000
la2m	1.009	1.056	.969	1.013	1.067	.965	1.000	1.000	1.000
hmoe	.974	.955	.993	.966	.937	.996	.997	.993	1.000
NFCorpus									
identity	.920	.459	1.136	.953	.386	1.153	.904	.544	1.039
relative	.852	1.218	.791	.898	1.254	.799	.907	1.248	.757
procrustes	.704	.517	.804	.817	.436	.874	.750	1.185	.551
cca	.918	1.294	.790	.959	1.237	.812	.980	1.401	.783
GW	.791	.519	.940	.888	.406	.989	.764	1.232	.556
mlp	.955	1.263	.841	.984	1.207	.871	.969	1.413	.759
vec2vec	.816	.000	1.177	.909	.000	1.199	.724	.000	1.048
embconv	.874	.410	1.104	.913	.376	1.130	.936	.886	.995
la2m	1.024	1.279	.940	1.025	1.268	.932	1.037	1.157	.981
hmoe	.976	.858	1.054	.979	.798	1.081	1.016	1.045	.991
ArguAna									
identity	.646	.178	1.119	.641	.130	1.168	.932	.863	1.001
relative	.983	1.339	.623	1.042	1.509	.560	.998	1.001	.994
procrustes	.598	.260	.939	.622	.206	1.052	.892	.890	.894
cca	.923	1.076	.768	.928	1.062	.791	.989	1.001	.977
GW	.570	.002	1.144	.618	.002	1.254	.492	.056	.929
mlp	.960	1.196	.721	.970	1.250	.681	.991	1.001	.981
vec2vec	.585	.000	1.176	.628	.000	1.277	.503	.004	1.001
embconv	.921	.854	.989	.930	.825	1.039	.997	.994	1.000
la2m	1.058	1.170	.945	1.078	1.230	.922	1.001	1.001	1.000
hmoe	1.019	1.108	.929	1.030	1.141	.914	1.001	1.001	1.000
FiQA									
identity	.738	.204	1.085	.797	.172	1.102	.826	.493	1.030
relative	.783	1.041	.678	.822	1.031	.672	.893	1.022	.812
procrustes	.630	.068	.995	.727	.050	1.030	.725	.509	.849
cca	.840	.871	.821	.856	.791	.840	.919	1.026	.853
GW	.647	.000	1.076	.746	.000	1.095	.597	.031	.969
mlp	.985	1.183	.863	.995	1.165	.863	.977	1.070	.913
vec2vec	.652	.000	1.088	.751	.000	1.105	.620	.000	1.030
embconv	.976	.984	.967	.963	.948	.963	1.000	1.022	.985
la2m	1.010	1.080	.966	1.015	1.093	.958	1.000	1.022	.988
hmoe	1.000	1.164	.898	1.000	1.165	.887	1.002	1.053	.968

Table A18: **bge-large→gte-large**, 1024 → 1024. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.537	.032	1.044	.545	.023	1.056	.662	.264	1.020
relative	.743	.945	.567	.755	.935	.574	.815	.980	.653
procrustes	.621	.560	.679	.603	.499	.699	.786	.919	.653
cca	.923	1.018	.824	.933	1.030	.825	.946	.993	.903
GW	.609	.490	.728	.597	.433	.758	.751	.899	.607
mlp	.883	1.092	.673	.908	1.110	.692	.826	1.000	.653
vec2vec	.431	.000	.853	.456	.000	.888	.400	.037	.723
embconv	.919	.840	1.006	.901	.795	1.015	.978	.960	.993
la2m	.988	.966	1.008	.982	.952	1.008	.997	.993	1.000
hmoe	.990	1.028	.952	.994	1.031	.953	.983	.993	.973
NFCorpus									
identity	.820	.000	1.177	.898	.000	1.190	.750	.046	1.097
relative	.671	1.196	.436	.795	1.196	.480	.625	1.276	.327
procrustes	.519	.598	.480	.680	.611	.559	.613	1.182	.327
cca	.885	1.270	.729	.953	1.186	.774	.866	1.349	.608
GW	.436	.188	.526	.609	.241	.632	.413	.752	.272
mlp	.787	1.579	.394	.899	1.511	.424	.647	1.442	.274
vec2vec	.553	.037	.823	.720	.045	.916	.342	.261	.358
embconv	.937	.817	.995	.955	.828	1.015	.988	1.060	.955
la2m	.992	.862	1.050	.992	.789	1.062	1.000	.981	.997
hmoe	.963	1.123	.904	.981	1.099	.926	.989	1.212	.866
ArguAna									
identity	.593	.002	1.181	.643	.002	1.286	.503	.004	1.004
relative	.841	1.271	.413	.861	1.367	.354	.934	1.003	.865
procrustes	.669	.455	.882	.678	.383	.974	.907	.973	.842
cca	.969	1.111	.827	.977	1.123	.829	.991	1.001	.981
GW	.524	.057	.991	.578	.043	1.113	.572	.337	.809
mlp	.927	1.131	.724	.934	1.145	.723	.938	1.003	.872
vec2vec	.552	.000	1.102	.608	.000	1.217	.458	.016	.902
embconv	1.020	1.064	.977	1.029	1.083	.974	1.000	1.000	1.000
la2m	.974	.896	1.052	.970	.870	1.070	.999	.997	1.000
hmoe	1.030	1.081	.979	1.042	1.105	.978	1.001	1.001	1.000
FiQA									
identity	.652	.000	1.100	.723	.000	1.103	.627	.003	1.037
relative	.685	1.006	.470	.712	.949	.484	.711	1.065	.489
procrustes	.559	.272	.756	.620	.212	.810	.632	.732	.553
cca	.911	1.065	.806	.938	1.028	.830	.920	1.100	.804
GW	.571	.079	.914	.651	.057	.954	.608	.522	.638
mlp	.934	1.170	.758	.976	1.145	.790	.846	1.118	.638
vec2vec	.651	.000	1.099	.722	.000	1.102	.592	.000	.990
embconv	.990	.955	1.007	.989	.950	1.016	1.003	1.020	.990
la2m	.964	.839	1.032	.967	.817	1.049	.979	.947	.995
hmoe	1.010	1.020	.991	1.010	1.013	1.002	.998	1.038	.972

Table A19: **bge-large→e5-large-v2**, 1024 → 1024. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.520	.000	1.044	.541	.000	1.051	.532	.000	1.020
relative	.575	.873	.288	.579	.862	.287	.688	.972	.428
procrustes	.287	.266	.309	.278	.228	.320	.604	.821	.401
cca	.739	.884	.597	.735	.870	.591	.838	.979	.706
GW	.374	.370	.382	.362	.318	.404	.567	.831	.325
mlp	.645	.989	.310	.659	.989	.311	.677	1.000	.373
vec2vec	.227	.006	.445	.245	.003	.472	.224	.117	.319
embconv	.783	.605	.973	.764	.543	.991	.942	.914	.973
la2m	.946	.895	.994	.942	.872	1.003	.979	.959	1.000
hmoe	.866	.991	.746	.871	.993	.745	.915	.993	.839
NFCorpus									
identity	.814	.000	1.155	.868	.000	1.178	.754	.000	1.095
relative	.455	1.363	.017	.588	1.344	.022	.422	1.374	.029
procrustes	.190	.502	.017	.253	.533	.023	.387	1.236	.027
cca	.565	1.240	.201	.650	1.236	.175	.630	1.455	.292
GW	.157	.366	.021	.200	.369	.026	.264	.959	.011
mlp	.569	1.712	.017	.690	1.561	.021	.481	1.627	.010
vec2vec	.063	.083	.017	.096	.094	.023	.150	.398	.007
embconv	.823	.842	.824	.867	.835	.852	.879	1.201	.784
la2m	.943	.822	1.022	.945	.780	1.060	.952	1.093	.923
hmoe	.765	1.418	.469	.811	1.343	.470	.787	1.509	.510
ArguAna									
identity	.585	.000	1.196	.626	.000	1.289	.500	.000	1.004
relative	.778	1.469	.057	.864	1.633	.049	.652	1.022	.280
procrustes	.267	.386	.142	.256	.351	.155	.526	.834	.216
cca	.744	1.198	.270	.748	1.226	.241	.826	1.024	.627
GW	.135	.006	.269	.158	.004	.321	.163	.120	.206
mlp	.718	1.332	.077	.753	1.397	.070	.627	1.025	.226
vec2vec	.585	.000	1.195	.625	.000	1.289	.501	.004	1.003
embconv	.944	1.002	.882	.943	1.003	.879	.995	1.002	.988
la2m	1.051	1.137	.963	1.051	1.142	.955	1.007	1.018	.996
hmoe	.869	1.165	.560	.865	1.197	.514	.948	1.010	.885
FiQA									
identity	.653	.000	1.113	.724	.000	1.099	.624	.000	1.070
relative	.383	.916	.055	.422	.864	.053	.478	1.053	.067
procrustes	.055	.053	.063	.057	.038	.069	.157	.299	.047
cca	.426	.911	.102	.469	.903	.098	.537	1.050	.155
GW	.086	.017	.124	.110	.013	.135	.092	.097	.081
mlp	.475	1.113	.052	.527	1.109	.049	.495	1.106	.051
vec2vec	.339	.000	.536	.419	.000	.582	.188	.010	.280
embconv	.864	.991	.786	.881	.994	.778	.919	1.004	.873
la2m	.973	1.007	.932	.985	1.003	.937	.976	1.030	.957
hmoe	.627	1.174	.251	.654	1.149	.220	.663	1.088	.338

Table A20: **gte-small→bge-small**, 384 → 384. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.795	.532	1.042	.779	.473	1.048	.956	.897	1.007
relative	.786	.805	.771	.768	.750	.780	.870	.969	.767
procrustes	.658	.457	.848	.634	.381	.858	.866	.931	.801
cca	.953	.987	.920	.951	.968	.926	.975	.993	.966
GW	.620	.319	.903	.606	.244	.925	.804	.856	.747
mlp	.937	.971	.880	.930	.936	.896	.899	1.021	.774
vec2vec	.493	.000	.973	.515	.000	.988	.527	.202	.822
embconv	1.003	1.016	.989	.995	1.000	.984	1.014	1.021	1.007
la2m	1.006	1.000	1.004	1.004	.988	1.006	1.007	1.007	1.007
hmoe	1.010	1.043	.975	1.003	1.028	.971	1.002	1.014	.993
NFCorpus									
identity	.911	.394	1.165	.940	.326	1.162	.933	.588	1.073
relative	.804	.949	.766	.899	.906	.796	.836	1.415	.523
procrustes	.706	.502	.816	.824	.473	.855	.783	1.133	.529
cca	.960	1.131	.903	.988	1.095	.878	1.006	1.269	.826
GW	.718	.167	.982	.835	.143	1.040	.596	.800	.463
mlp	.937	1.129	.860	.956	.928	.900	.824	1.417	.469
vec2vec	.807	.002	1.186	.899	.003	1.200	.660	.176	.901
embconv	.993	1.043	.979	.986	1.042	.966	1.005	1.041	.991
la2m	.988	.980	1.000	.983	.980	.984	1.018	1.018	1.018
hmoe	.993	1.067	.966	.986	1.060	.948	1.007	1.068	.981
ArguAna									
identity	.627	.105	1.158	.655	.075	1.249	.880	.757	1.004
relative	.825	1.093	.553	.827	1.095	.554	.944	1.006	.882
procrustes	.676	.438	.918	.686	.365	1.015	.927	.961	.892
cca	.976	1.081	.869	.971	1.084	.855	.996	1.004	.987
GW	.546	.047	1.052	.595	.035	1.168	.619	.375	.864
mlp	.897	.846	.949	.879	.747	1.014	.950	1.004	.896
vec2vec	.585	.000	1.179	.637	.000	1.289	.578	.155	1.003
embconv	.996	1.012	.979	.995	1.017	.973	1.002	1.004	1.000
la2m	.995	.995	.994	.992	.991	.994	1.001	1.001	1.000
hmoe	1.001	1.031	.970	.999	1.034	.962	1.001	1.004	.999
FiQA									
identity	.761	.286	1.084	.781	.245	1.078	.875	.618	1.013
relative	.748	.817	.697	.774	.723	.755	.840	1.116	.639
procrustes	.602	.157	.900	.673	.117	.947	.732	.687	.730
cca	.913	.943	.886	.921	.915	.909	.988	1.112	.918
GW	.634	.040	1.049	.703	.027	1.060	.728	.525	.831
mlp	.891	.813	.929	.891	.710	.968	.964	1.161	.831
vec2vec	.640	.000	1.091	.707	.000	1.084	.619	.000	1.020
embconv	.979	1.024	.952	.976	1.022	.947	1.006	1.028	.992
la2m	.982	.992	.983	.982	.986	.978	1.000	.991	1.002
hmoe	.985	1.060	.936	.989	1.066	.940	1.006	1.058	.978

Table A21: **qwen3-8b→nemotron**, 4096 → 4096. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.535	.000	1.050	.561	.000	1.060	.535	.000	1.007
relative	.700	.892	.531	.693	.871	.524	.868	.979	.773
procrustes	.477	.378	.564	.462	.310	.580	.750	.874	.640
cca	.920	1.021	.834	.926	1.012	.845	.973	1.014	.940
GW	.539	.554	.513	.543	.502	.547	.701	.924	.507
mlp	.934	1.134	.754	.960	1.163	.759	.929	1.021	.847
vec2vec	.535	.000	1.050	.561	.000	1.060	.535	.000	1.007
embconv	.735	.452	1.016	.712	.383	1.028	.933	.881	.987
la2m	.945	.854	1.032	.935	.821	1.042	1.000	1.000	1.000
hmoe	.939	.873	1.002	.927	.830	1.009	.998	1.010	.987
NFCorpus									
identity	.797	.000	1.166	.885	.000	1.191	.759	.000	1.259
relative	1.114	2.170	.612	1.134	2.346	.619	.892	1.775	.483
procrustes	.640	1.019	.369	.758	1.077	.472	.708	1.524	.272
cca	1.057	1.348	.972	.997	1.204	.946	.952	1.617	.660
GW	.606	.989	.285	.727	.930	.389	.497	1.283	.207
mlp	1.097	1.679	.834	1.043	1.565	.856	.906	1.788	.516
vec2vec	.653	.152	.837	.786	.159	.917	.253	.545	.181
embconv	.885	.385	1.143	.916	.375	1.157	.950	1.012	.966
la2m	.909	.465	1.121	.928	.465	1.154	1.069	1.010	1.171
hmoe	.929	.540	1.106	.942	.493	1.144	1.006	1.132	.953
ArguAna									
identity	.578	.000	1.166	.627	.000	1.267	.500	.000	1.000
relative	.988	1.256	.715	1.033	1.389	.669	.998	1.000	.996
procrustes	.664	.231	1.105	.678	.163	1.205	.954	.944	.963
cca	.974	.973	.974	.976	.935	1.017	.998	1.000	.996
GW	.579	.016	1.152	.627	.011	1.256	.742	.506	.979
mlp	1.000	1.062	.938	1.003	1.058	.946	.998	1.000	.996
vec2vec	.578	.000	1.166	.627	.000	1.267	.500	.000	1.000
embconv	.972	.927	1.018	.991	.915	1.069	1.000	1.000	1.000
la2m	1.063	1.141	.984	1.087	1.187	.985	1.000	1.000	1.000
hmoe	1.034	1.105	.963	1.049	1.134	.962	1.000	1.000	1.000
FiQA									
identity	.659	.000	1.177	.752	.000	1.198	.745	.000	1.427
relative	.775	1.384	.311	.814	1.425	.358	1.247	2.111	.416
procrustes	.552	.114	.701	.621	.065	.845	.523	.597	.416
cca	.883	1.140	.606	.932	1.048	.700	1.019	1.540	.584
GW	.619	.000	1.136	.709	.000	1.164	.342	.039	.619
mlp	1.108	1.768	.515	1.095	1.733	.565	1.195	1.910	.549
vec2vec	.659	.000	1.177	.752	.000	1.198	.745	.000	1.427
embconv	.735	.613	.868	.776	.588	.913	.935	.930	.936
la2m	1.005	.847	1.058	1.009	.829	1.089	1.134	1.037	1.267
hmoe	.897	1.152	.672	.914	1.059	.766	.943	1.224	.794

Table A22: **gte→nv-embed-v2**, 3584 → 4096. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.520	.000	1.039	.532	.000	1.051	.539	.000	1.006
relative	.879	.832	.950	.852	.771	.950	.978	.993	.968
procrustes	.630	.238	1.019	.602	.161	1.037	.922	.901	.949
cca	.915	.824	1.010	.894	.776	1.025	1.000	1.007	.994
GW	.603	.183	1.021	.577	.114	1.034	.932	.912	.949
mlp	.945	.889	1.002	.928	.845	1.017	.995	1.007	.987
vec2vec	.520	.000	1.039	.532	.000	1.051	.539	.000	1.006
embconv	.698	.375	1.025	.657	.269	1.040	.950	.895	1.000
la2m	.954	.892	1.017	.935	.850	1.024	1.003	1.007	1.000
hmoe	.945	.872	1.022	.923	.824	1.031	1.003	1.007	1.000
NFCorpus									
identity	.814	.000	1.172	.871	.000	1.155	.733	.000	1.075
relative	.905	.933	.926	.914	.903	.929	.955	1.121	.917
procrustes	.775	.230	1.019	.869	.222	1.032	.801	.866	.768
cca	.899	.823	.936	.899	.763	.950	.967	1.102	.931
GW	.804	.077	1.107	.862	.057	1.124	.811	.702	.853
mlp	.923	.883	.928	.916	.799	.940	.970	1.176	.898
vec2vec	.814	.000	1.172	.871	.000	1.155	.733	.000	1.075
embconv	.904	.568	1.057	.901	.557	1.054	.947	.936	.999
la2m	.958	.704	1.064	.933	.670	1.062	.974	.954	1.007
hmoe	.948	.742	1.038	.922	.726	1.035	.970	.973	.994
ArguAna									
identity	.578	.000	1.154	.624	.000	1.253	.501	.000	1.000
relative	.945	1.167	.724	.963	1.245	.677	.999	1.001	.997
procrustes	.616	.108	1.124	.647	.076	1.224	.935	.886	.984
cca	.922	.809	1.036	.907	.710	1.106	.996	1.001	.991
GW	.634	.131	1.136	.660	.089	1.237	.978	.974	.982
mlp	.974	.985	.962	.966	.948	.985	.997	1.001	.993
vec2vec	.578	.000	1.154	.624	.000	1.253	.501	.000	1.000
embconv	.921	.818	1.023	.925	.768	1.083	.999	.997	1.000
la2m	.949	.881	1.016	.933	.830	1.037	1.001	1.001	1.000
hmoe	.976	.962	.990	.968	.932	1.004	1.001	1.001	1.000
FiQA									
identity	.647	.000	1.090	.736	.000	1.104	.623	.000	1.034
relative	.812	1.097	.697	.855	1.087	.696	.917	1.061	.815
procrustes	.589	.073	.943	.681	.054	.982	.692	.498	.805
cca	.829	.778	.880	.841	.689	.911	.939	1.019	.896
GW	.622	.009	1.031	.718	.005	1.063	.605	.230	.843
mlp	.940	1.115	.845	.960	1.102	.845	.999	1.089	.942
vec2vec	.647	.000	1.090	.736	.000	1.104	.623	.000	1.034
embconv	.908	.811	.993	.911	.763	1.000	.986	.981	.993
la2m	.934	.771	1.044	.932	.732	1.053	.995	.981	1.013
hmoe	.955	.991	.958	.965	.973	.969	1.005	1.050	.976

Table A23: **gte→qwen3-8b**, 3584 → 4096. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.522	.000	1.035	.533	.000	1.043	.540	.000	1.020
relative	.891	.911	.888	.879	.869	.897	.969	.993	.948
procrustes	.686	.407	.954	.649	.301	.972	.947	.969	.928
cca	.948	.934	.962	.932	.895	.970	.993	1.007	.980
GW	.628	.257	.986	.603	.176	1.006	.940	.966	.922
mlp	.955	.946	.959	.945	.912	.967	.986	1.007	.967
vec2vec	.522	.000	1.035	.533	.000	1.043	.540	.000	1.020
embconv	.807	.601	1.012	.757	.490	1.022	.988	.983	.994
la2m	.985	.957	1.011	.977	.930	1.017	1.000	1.007	.994
hmoe	.983	.957	1.009	.971	.926	1.015	1.000	1.007	.994
NFCorpus									
identity	.820	.000	1.156	.912	.000	1.171	.740	.000	1.091
relative	.911	.961	.911	.934	.885	.923	.913	1.244	.795
procrustes	.822	.361	1.005	.927	.334	1.037	.820	1.093	.694
cca	.957	.988	.964	.971	.881	.984	1.006	1.251	.895
GW	.829	.185	1.080	.908	.142	1.123	.800	1.010	.698
mlp	.950	.899	.969	.960	.781	.995	.982	1.272	.843
vec2vec	.820	.000	1.156	.912	.000	1.171	.740	.000	1.091
embconv	.901	.547	1.061	.949	.471	1.094	.984	.989	.989
la2m	.974	.831	1.042	.973	.770	1.051	1.014	1.062	.997
hmoe	.966	.841	1.029	.962	.774	1.045	1.008	1.073	.976
ArguAna									
identity	.567	.000	1.149	.616	.000	1.253	.500	.000	1.000
relative	.905	.991	.816	.899	1.010	.784	.999	1.000	.999
procrustes	.641	.190	1.105	.665	.140	1.208	.936	.877	.994
cca	.938	.848	1.031	.931	.775	1.092	.999	1.000	.997
GW	.584	.044	1.138	.625	.026	1.244	.923	.853	.993
mlp	.963	.937	.989	.954	.902	1.007	.999	1.000	.997
vec2vec	.567	.000	1.149	.616	.000	1.253	.507	.014	.999
embconv	.924	.780	1.073	.926	.722	1.138	.996	.994	.999
la2m	.934	.834	1.037	.917	.774	1.065	1.000	1.000	1.000
hmoe	.971	.924	1.020	.963	.888	1.041	.999	.999	.999
FiQA									
identity	.658	.000	1.091	.750	.000	1.101	.612	.000	1.007
relative	.820	.991	.746	.834	.946	.750	.939	1.034	.874
procrustes	.674	.102	1.044	.757	.074	1.067	.853	.746	.912
cca	.873	.799	.917	.882	.705	.947	.962	1.033	.903
GW	.657	.010	1.084	.749	.006	1.097	.732	.387	.951
mlp	.973	1.016	.943	.981	.987	.954	.989	1.052	.946
vec2vec	.658	.000	1.091	.750	.000	1.101	.612	.000	1.007
embconv	.968	.924	.997	.976	.886	1.006	.998	1.023	.989
la2m	.946	.774	1.055	.950	.719	1.069	.997	.988	1.002
hmoe	.981	.988	.974	.986	.963	.978	.996	1.038	.967

Table A24: **kalm→qwen3-8b**, 3840 → 4096. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.522	.000	1.035	.533	.000	1.043	.540	.000	1.020
relative	.892	.912	.896	.883	.877	.908	.978	.997	.961
procrustes	.689	.403	.966	.657	.309	.984	.949	.966	.935
cca	.937	.905	.964	.921	.865	.972	.993	1.007	.980
GW	.664	.325	.990	.636	.239	1.008	.944	.973	.922
mlp	.959	.948	.970	.952	.916	.981	.990	1.007	.974
vec2vec	.522	.000	1.035	.533	.000	1.043	.540	.000	1.020
embconv	.767	.512	1.024	.726	.417	1.034	.980	.966	.994
la2m	.954	.888	1.016	.938	.849	1.022	1.000	1.000	1.000
hmoe	.982	.961	1.002	.975	.939	1.009	1.000	1.007	.994
NFCorpus									
identity	.820	.000	1.156	.912	.000	1.171	.740	.000	1.091
relative	.912	1.004	.910	.938	.975	.933	.924	1.228	.806
procrustes	.823	.333	1.019	.925	.296	1.046	.820	1.070	.690
cca	.940	.912	.961	.965	.834	.993	.979	1.189	.891
GW	.839	.257	1.068	.913	.178	1.109	.791	1.010	.679
mlp	.943	.877	.980	.965	.779	1.005	.978	1.231	.860
vec2vec	.820	.000	1.156	.912	.000	1.171	.740	.000	1.091
embconv	.889	.468	1.081	.936	.404	1.104	.942	.893	.992
la2m	.966	.737	1.065	.962	.705	1.073	.997	.968	1.016
hmoe	.961	.834	1.030	.962	.796	1.043	.998	1.058	.980
ArguAna									
identity	.567	.000	1.149	.616	.000	1.253	.500	.000	1.000
relative	.889	.999	.776	.883	1.017	.743	.999	1.000	.999
procrustes	.629	.166	1.105	.652	.117	1.206	.952	.909	.994
cca	.943	.878	1.010	.933	.819	1.052	.999	1.000	.997
GW	.615	.112	1.132	.646	.074	1.238	.969	.947	.990
mlp	.954	.935	.972	.940	.905	.976	.999	1.000	.997
vec2vec	.567	.000	1.149	.616	.000	1.253	.500	.000	1.000
embconv	.937	.816	1.061	.938	.761	1.120	.997	.996	.999
la2m	.951	.901	1.002	.935	.870	1.002	.999	.999	1.000
hmoe	.964	.940	.988	.950	.919	.983	.999	1.000	.999
FiQA									
identity	.658	.000	1.091	.750	.000	1.101	.612	.000	1.007
relative	.818	.995	.740	.855	.975	.753	.933	1.022	.862
procrustes	.680	.112	1.048	.757	.074	1.071	.860	.777	.908
cca	.883	.820	.922	.899	.732	.958	.962	1.031	.906
GW	.653	.001	1.082	.747	.001	1.095	.701	.311	.952
mlp	.959	1.008	.930	.971	.980	.940	.987	1.055	.941
vec2vec	.658	.000	1.091	.750	.000	1.101	.612	.000	1.007
embconv	.961	.947	.973	.970	.933	.976	.995	1.014	.985
la2m	.965	.894	1.019	.969	.873	1.030	1.000	1.003	.998
hmoe	.967	1.024	.941	.974	1.010	.934	.999	1.049	.968

Table A25: **bge-large→nv-embed-v2**, 1024 → 4096. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.520	.000	1.039	.532	.000	1.051	.539	.000	1.006
relative	.818	.698	.954	.770	.608	.948	.973	.990	.962
procrustes	.588	.158	1.018	.570	.103	1.033	.890	.837	.949
cca	.841	.662	1.017	.800	.572	1.032	.988	.986	.994
GW	.585	.137	1.030	.573	.091	1.044	.854	.753	.943
mlp	.867	.724	1.007	.831	.642	1.022	.985	.993	.981
vec2vec	.520	.000	1.039	.532	.000	1.051	.542	.014	1.000
embconv	.659	.293	1.030	.622	.201	1.044	.956	.912	1.000
la2m	.868	.702	1.030	.829	.618	1.038	.997	.993	1.000
hmoe	.877	.733	1.023	.838	.650	1.032	.997	.993	1.000
NFCorpus									
identity	.814	.000	1.172	.871	.000	1.155	.733	.000	1.075
relative	.867	.641	1.002	.897	.597	1.002	.911	1.089	.872
procrustes	.785	.258	1.015	.872	.241	1.029	.814	.855	.791
cca	.878	.598	1.006	.909	.557	1.018	.947	1.053	.930
GW	.802	.044	1.126	.860	.041	1.126	.766	.573	.863
mlp	.884	.665	.981	.912	.589	.999	.945	1.120	.882
vec2vec	.814	.000	1.172	.871	.000	1.155	.733	.000	1.075
embconv	.866	.364	1.104	.889	.340	1.107	.912	.781	1.006
la2m	.907	.402	1.127	.902	.351	1.128	.959	.867	1.014
hmoe	.928	.606	1.072	.924	.565	1.077	.972	.989	.978
ArguAna									
identity	.578	.000	1.154	.624	.000	1.253	.501	.000	1.000
relative	.917	1.126	.709	.923	1.177	.667	.995	1.001	.989
procrustes	.627	.127	1.126	.656	.093	1.225	.958	.931	.984
cca	.832	.555	1.109	.814	.432	1.200	.996	1.000	.993
GW	.575	.000	1.149	.621	.000	1.248	.524	.063	.984
mlp	.907	.753	1.059	.888	.648	1.131	.996	1.001	.991
vec2vec	.578	.000	1.154	.624	.000	1.253	.504	.007	1.000
embconv	.841	.586	1.094	.836	.497	1.179	.996	.993	1.000
la2m	.899	.720	1.077	.888	.645	1.132	1.000	1.000	1.000
hmoe	.946	.839	1.051	.940	.777	1.105	1.001	1.001	1.000
FiQA									
identity	.647	.000	1.090	.736	.000	1.104	.623	.000	1.034
relative	.713	.727	.715	.756	.673	.746	.850	.971	.768
procrustes	.577	.022	.955	.678	.016	.999	.633	.339	.818
cca	.742	.466	.939	.761	.379	.970	.886	.863	.894
GW	.626	.000	1.045	.723	.000	1.073	.540	.013	.887
mlp	.804	.773	.822	.829	.736	.849	.905	1.009	.830
vec2vec	.647	.000	1.090	.736	.000	1.104	.623	.000	1.034
embconv	.765	.342	1.052	.787	.270	1.074	.900	.750	1.002
la2m	.762	.301	1.075	.783	.234	1.087	.898	.709	1.022
hmoe	.824	.629	.966	.835	.577	.978	.943	.925	.962

Table A26: **nv-embed-v2→bge-large**, 4096 → 1024. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.526	.000	1.045	.539	.000	1.047	.529	.000	1.020
relative	.793	.825	.785	.778	.783	.797	.938	.973	.905
procrustes	.591	.356	.816	.580	.279	.855	.852	.912	.792
cca	.890	.866	.919	.877	.823	.940	.979	.986	.973
GW	.589	.310	.852	.583	.244	.888	.807	.855	.752
mlp	.897	.971	.829	.898	.955	.846	.959	1.000	.922
vec2vec	.526	.000	1.045	.539	.000	1.047	.522	.000	1.007
embconv	.748	.497	1.001	.712	.406	1.013	.939	.888	.993
la2m	.863	.697	1.030	.820	.609	1.033	.997	.993	1.000
hmoe	.941	.884	.999	.930	.851	1.010	.997	1.000	.993
NFCorpus									
identity	.826	.000	1.172	.902	.000	1.150	.738	.000	1.087
relative	.831	1.232	.729	.891	1.266	.731	.836	1.229	.665
procrustes	.673	.458	.786	.817	.386	.849	.685	1.133	.477
cca	.907	1.060	.855	.954	1.014	.898	.942	1.327	.786
GW	.716	.317	.913	.867	.281	.980	.672	1.127	.470
mlp	.905	1.195	.794	.955	1.113	.828	.888	1.406	.661
vec2vec	.826	.000	1.172	.902	.000	1.150	.738	.003	1.083
embconv	.836	.345	1.034	.896	.309	1.069	.923	.911	.931
la2m	.927	.596	1.089	.927	.527	1.077	.991	.964	1.033
hmoe	.939	.818	1.002	.949	.797	1.021	.990	1.106	.945
ArguAna									
identity	.583	.000	1.168	.639	.000	1.280	.502	.000	1.004
relative	.906	1.173	.639	.934	1.271	.595	.984	.994	.973
procrustes	.603	.361	.845	.620	.306	.935	.895	.924	.866
cca	.934	.898	.971	.939	.859	1.019	.994	1.001	.986
GW	.527	.000	1.056	.587	.000	1.176	.475	.092	.859
mlp	.923	1.148	.697	.928	1.192	.662	.976	1.004	.947
vec2vec	.583	.000	1.168	.639	.000	1.280	.502	.003	1.003
embconv	.970	.957	.983	.988	.952	1.025	.996	.997	.996
la2m	.979	.897	1.061	.983	.856	1.111	1.001	1.001	1.000
hmoe	1.008	1.066	.951	1.020	1.093	.945	.998	1.000	.996
FiQA									
identity	.647	.000	1.101	.716	.000	1.105	.611	.000	1.018
relative	.782	1.050	.662	.793	.960	.696	.833	1.100	.675
procrustes	.572	.118	.872	.643	.086	.926	.676	.613	.725
cca	.904	1.149	.751	.922	1.070	.780	.923	1.144	.774
GW	.632	.000	1.074	.705	.000	1.089	.530	.038	.852
mlp	.989	1.329	.766	.999	1.261	.781	.926	1.159	.782
vec2vec	.647	.000	1.101	.716	.000	1.105	.611	.000	1.018
embconv	.987	1.070	.948	.972	1.022	.949	.996	1.047	.978
la2m	.915	.718	1.063	.878	.610	1.067	1.011	1.026	1.008
hmoe	.992	1.223	.863	.981	1.169	.865	.993	1.121	.922

Table A27: **qwen3-0.6b→bge-large**, 1024 → 1024. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.526	.000	1.045	.539	.000	1.047	.529	.000	1.020
relative	.801	.875	.752	.792	.849	.757	.900	.959	.840
procrustes	.597	.370	.815	.586	.296	.849	.850	.905	.792
cca	.860	.819	.901	.847	.767	.924	.977	.983	.973
GW	.549	.225	.860	.546	.161	.898	.798	.832	.758
mlp	.868	.907	.829	.867	.869	.858	.907	.993	.820
vec2vec	.473	.004	.929	.498	.003	.957	.430	.034	.786
embconv	.871	.739	1.009	.834	.662	1.017	.984	.976	.993
la2m	.857	.688	1.029	.823	.617	1.035	.993	.986	1.000
hmoe	.919	.832	1.010	.899	.786	1.020	.997	1.000	.993
NFCorpus									
identity	.826	.000	1.172	.902	.000	1.150	.738	.000	1.087
relative	.811	1.137	.736	.897	1.165	.768	.797	1.292	.593
procrustes	.659	.392	.783	.811	.386	.842	.673	1.083	.517
cca	.863	.656	.945	.929	.632	.975	.900	1.188	.784
GW	.667	.070	.943	.831	.082	1.012	.494	.536	.474
mlp	.844	1.005	.777	.922	.904	.835	.807	1.415	.547
vec2vec	.826	.000	1.172	.902	.000	1.150	.727	.016	1.062
embconv	.892	.616	1.026	.927	.614	1.026	.953	.963	.957
la2m	.915	.515	1.090	.929	.449	1.086	.977	.931	1.016
hmoe	.910	.620	1.035	.926	.562	1.057	.976	1.052	.942
ArguAna									
identity	.583	.000	1.168	.639	.000	1.280	.502	.000	1.004
relative	.874	1.185	.561	.891	1.277	.503	.973	.997	.948
procrustes	.625	.398	.853	.632	.324	.940	.908	.944	.872
cca	.943	1.104	.781	.939	1.118	.760	.988	1.006	.970
GW	.523	.070	.979	.573	.052	1.096	.687	.537	.836
mlp	.894	1.117	.670	.895	1.133	.656	.956	1.006	.906
vec2vec	.564	.001	1.129	.622	.001	1.245	.475	.039	.912
embconv	.991	1.072	.910	.994	1.086	.900	.999	1.004	.994
la2m	1.016	1.058	.974	1.028	1.081	.976	1.002	1.006	.999
hmoe	1.008	1.100	.915	1.011	1.125	.898	1.000	1.006	.994
FiQA									
identity	.647	.000	1.101	.716	.000	1.105	.611	.000	1.018
relative	.731	.953	.598	.764	.900	.641	.802	1.055	.636
procrustes	.567	.123	.858	.633	.090	.909	.682	.625	.694
cca	.852	.982	.750	.879	.919	.788	.907	1.109	.766
GW	.621	.009	1.051	.698	.005	1.067	.557	.177	.809
mlp	.905	1.026	.802	.923	.958	.840	.894	1.123	.748
vec2vec	.647	.000	1.101	.716	.000	1.105	.593	.000	.984
embconv	.912	.971	.876	.917	.926	.886	.987	1.042	.964
la2m	.905	.708	1.036	.901	.655	1.054	.984	.954	1.000
hmoe	.911	1.016	.842	.922	.978	.862	.981	1.099	.902

Table A28: **nemotron→bge-small**, 4096 → 384. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.526	.000	1.042	.544	.000	1.049	.526	.000	1.007
relative	.777	.790	.777	.755	.747	.769	.917	.986	.846
procrustes	.603	.339	.851	.583	.270	.863	.851	.893	.808
cca	.863	.748	.970	.835	.672	.984	.974	.972	.980
GW	.495	.066	.910	.510	.043	.934	.643	.538	.740
mlp	.892	.922	.864	.878	.883	.867	.937	1.007	.866
vec2vec	.526	.000	1.042	.544	.000	1.049	.526	.000	1.007
embconv	.493	.000	.974	.516	.000	.989	.506	.103	.870
la2m	.778	.513	1.034	.734	.404	1.044	.979	.948	1.007
hmoe	.930	.876	.988	.917	.845	.993	.988	.986	.993
NFCorpus									
identity	.817	.000	1.198	.899	.000	1.204	.727	.000	1.091
relative	.850	1.102	.802	.894	1.136	.804	.878	1.262	.662
procrustes	.678	.404	.838	.802	.379	.870	.698	1.030	.499
cca	.903	.837	.959	.935	.795	.966	.971	1.230	.843
GW	.715	.188	.954	.869	.212	1.023	.640	.936	.474
mlp	.874	1.029	.830	.928	.954	.850	.915	1.377	.649
vec2vec	.817	.000	1.198	.899	.000	1.204	.726	.043	1.055
embconv	.785	.017	1.132	.888	.015	1.165	.676	.415	.757
la2m	.898	.384	1.148	.920	.311	1.177	.950	.834	1.038
hmoe	.931	.753	1.032	.963	.732	1.054	.977	1.069	.954
ArguAna									
identity	.586	.000	1.181	.638	.000	1.290	.501	.000	1.004
relative	.911	1.186	.631	.934	1.269	.591	.976	.997	.955
procrustes	.560	.173	.954	.589	.132	1.056	.827	.753	.902
cca	.906	.726	1.090	.899	.636	1.167	.995	.991	.999
GW	.530	.001	1.069	.585	.001	1.184	.479	.072	.889
mlp	.950	1.134	.763	.953	1.146	.754	.974	1.006	.942
vec2vec	.585	.000	1.180	.637	.000	1.290	.499	.000	1.000
embconv	.911	.773	1.051	.923	.719	1.132	.988	.977	.999
la2m	.868	.605	1.137	.849	.480	1.226	1.001	1.003	1.000
hmoe	1.026	1.096	.954	1.032	1.111	.952	1.001	1.004	.999
FiQA									
identity	.640	.000	1.091	.707	.000	1.084	.621	.000	1.024
relative	.782	.892	.723	.807	.816	.778	.876	1.132	.692
procrustes	.587	.096	.919	.666	.068	.962	.724	.600	.774
cca	.865	.911	.839	.884	.866	.872	.993	1.204	.864
GW	.634	.000	1.080	.704	.000	1.076	.549	.003	.893
mlp	.918	1.084	.802	.929	1.039	.831	.980	1.212	.832
vec2vec	.640	.000	1.091	.707	.000	1.084	.621	.000	1.024
embconv	.887	.801	.963	.892	.763	.967	1.005	1.062	.977
la2m	.775	.351	1.079	.777	.265	1.075	.975	.921	1.011
hmoe	.933	1.011	.884	.936	.976	.893	1.025	1.185	.932

Table A29: **openai→gte**, 1536 → 3584. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.519	.000	1.029	.527	.000	1.034	.535	.000	1.000
relative	.859	.766	.964	.822	.691	.965	.973	.987	.962
procrustes	.642	.265	1.009	.610	.181	1.021	.941	.936	.949
cca	.889	.764	1.006	.856	.692	1.013	.990	1.000	.981
GW	.593	.148	1.026	.574	.096	1.032	.948	.939	.962
mlp	.901	.799	.998	.874	.736	1.006	.983	.993	.975
vec2vec	.519	.000	1.029	.527	.000	1.034	.535	.000	1.000
embconv	.819	.619	1.016	.775	.528	1.023	.982	.963	1.000
la2m	.893	.759	1.017	.864	.703	1.020	.997	.993	1.000
hmoe	.918	.828	1.006	.892	.774	1.010	1.000	1.000	1.000
NFCorpus									
identity	.826	.000	1.184	.879	.000	1.158	.726	.000	1.078
relative	.915	.749	1.007	.913	.655	1.046	.971	1.167	.891
procrustes	.827	.204	1.083	.888	.183	1.104	.877	.931	.844
cca	.931	.681	1.029	.928	.550	1.046	.985	1.100	.936
GW	.828	.021	1.167	.880	.014	1.149	.824	.632	.901
mlp	.928	.636	1.057	.929	.512	1.077	.992	1.147	.928
vec2vec	.826	.000	1.184	.879	.000	1.158	.725	.000	1.076
embconv	.887	.515	1.063	.897	.446	1.079	.970	.957	.997
la2m	.941	.587	1.114	.922	.486	1.106	.987	.938	1.021
hmoe	.942	.691	1.058	.931	.583	1.057	.988	1.015	.994
ArguAna									
identity	.586	.000	1.178	.634	.000	1.278	.500	.000	1.000
relative	.954	1.167	.738	.989	1.254	.719	.998	1.000	.996
procrustes	.680	.246	1.118	.700	.180	1.229	.986	.983	.989
cca	.882	.690	1.076	.877	.589	1.171	.999	1.000	.997
GW	.656	.185	1.132	.682	.131	1.242	.976	.981	.972
mlp	.970	.985	.956	.972	.953	.991	.999	1.000	.997
vec2vec	.586	.000	1.178	.634	.000	1.278	.500	.000	1.000
embconv	.936	.815	1.058	.947	.760	1.137	1.000	1.000	1.000
la2m	.969	.910	1.028	.957	.866	1.050	1.000	1.000	1.000
hmoe	.979	.952	1.007	.979	.919	1.040	1.000	1.000	1.000
FiQA									
identity	.651	.000	1.089	.741	.000	1.093	.616	.000	1.024
relative	.706	.791	.682	.756	.783	.689	.891	.952	.835
procrustes	.645	.065	1.030	.730	.042	1.056	.790	.622	.884
cca	.840	.691	.924	.859	.624	.950	.936	.982	.895
GW	.646	.011	1.066	.738	.006	1.081	.680	.319	.911
mlp	.891	.878	.891	.901	.830	.902	.946	.996	.909
vec2vec	.651	.000	1.089	.741	.000	1.093	.616	.000	1.024
embconv	.854	.633	.991	.865	.573	1.002	.971	.940	.990
la2m	.851	.526	1.059	.866	.465	1.069	.975	.912	1.012
hmoe	.900	.813	.947	.905	.760	.956	.969	.975	.962

Table A30: **openai→bge-large**, 1536 → 1024. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.526	.000	1.045	.539	.000	1.047	.529	.000	1.020
relative	.814	.886	.769	.816	.867	.783	.904	.966	.840
procrustes	.629	.437	.808	.620	.365	.845	.869	.925	.813
cca	.821	.714	.923	.793	.636	.940	.972	.973	.973
GW	.579	.288	.852	.579	.234	.890	.810	.871	.745
mlp	.876	.934	.812	.876	.904	.837	.921	.993	.847
vec2vec	.526	.000	1.045	.539	.000	1.047	.522	.000	1.007
embconv	.804	.603	1.010	.777	.537	1.018	.972	.952	.993
la2m	.895	.759	1.028	.864	.693	1.030	.993	.993	.993
hmoe	.934	.867	.999	.917	.827	1.005	.997	1.000	.993
NFCorpus									
identity	.826	.000	1.172	.902	.000	1.150	.738	.000	1.087
relative	.825	1.193	.723	.911	1.224	.739	.790	1.264	.595
procrustes	.683	.452	.797	.847	.468	.861	.701	1.097	.525
cca	.901	.843	.917	.953	.805	.961	.929	1.234	.796
GW	.715	.269	.921	.861	.253	.994	.655	1.016	.488
mlp	.891	1.165	.772	.952	1.074	.820	.811	1.405	.559
vec2vec	.826	.000	1.172	.902	.000	1.150	.746	.012	1.086
embconv	.892	.586	1.040	.929	.577	1.057	.979	1.051	.951
la2m	.929	.569	1.087	.937	.528	1.087	.999	.993	1.019
hmoe	.953	.785	1.030	.963	.752	1.037	.981	1.101	.938
ArguAna									
identity	.583	.000	1.168	.639	.000	1.280	.502	.000	1.004
relative	.840	1.139	.539	.853	1.211	.494	.974	.997	.951
procrustes	.640	.444	.836	.648	.371	.926	.911	.961	.860
cca	.925	1.039	.810	.915	1.035	.795	.991	1.006	.977
GW	.502	.002	1.005	.562	.001	1.124	.429	.014	.845
mlp	.885	1.110	.659	.883	1.128	.637	.963	1.006	.921
vec2vec	.582	.000	1.167	.638	.000	1.279	.501	.006	.997
embconv	.980	1.011	.948	.984	1.010	.958	1.000	1.004	.996
la2m	.985	1.000	.970	.981	.992	.970	1.001	1.003	1.000
hmoe	.994	1.088	.900	.992	1.106	.876	1.000	1.006	.994
FiQA									
identity	.647	.000	1.101	.716	.000	1.105	.611	.000	1.018
relative	.720	.890	.638	.754	.829	.680	.800	1.059	.622
procrustes	.571	.162	.834	.640	.133	.883	.667	.621	.677
cca	.855	.962	.783	.876	.895	.821	.917	1.091	.804
GW	.620	.025	1.039	.699	.018	1.062	.618	.321	.797
mlp	.899	1.077	.774	.915	1.021	.797	.913	1.123	.775
vec2vec	.647	.000	1.101	.716	.000	1.105	.611	.000	1.018
embconv	.932	.936	.948	.918	.882	.957	.995	1.034	.973
la2m	.899	.687	1.048	.891	.631	1.067	.987	.952	1.004
hmoe	.951	1.089	.871	.955	1.057	.881	.983	1.087	.919

Table A31: **mistral→openai**, 1024 → 1536. Each entry reports quality retention: the mixed index's score normalised by the score of an index built independently by the target model. A value of 1 therefore indicates parity with the target's native indexing quality. On the full index and on D1, values above 1 indicate that the translated index performs even better than the target model's native index; on D2, where parity is the ideal rather than the floor, they indicate the opposite, that the translated rows are failing to compete for ranks the native rows then keep. This is why identity, which translates nothing, stands above 1 on D2 in every corpus.

	nDCG@10 (ret.)			MRR@10 (ret.)			Recall@100 (ret.)		
mapper	full	D1	D2	full	D1	D2	full	D1	D2
SciFact									
identity	.516	.000	1.033	.526	.000	1.032	.533	.000	1.006
relative	.833	.818	.865	.804	.741	.882	.926	.993	.863
procrustes	.583	.269	.900	.553	.189	.903	.892	.912	.882
cca	.876	.782	.972	.838	.708	.965	.952	.980	.928
GW	.539	.103	.973	.538	.074	.981	.796	.723	.863
mlp	.896	.840	.951	.865	.765	.959	.944	1.007	.889
vec2vec	.516	.000	1.033	.526	.000	1.032	.533	.000	1.006
embconv	.797	.608	.998	.749	.508	1.000	.969	.952	.987
la2m	.945	.880	1.012	.920	.833	1.014	.997	.993	1.000
hmoe	.971	.971	.980	.957	.949	.975	.986	.993	.980
NFCorpus									
identity	.832	.000	1.186	.914	.000	1.189	.757	.000	1.097
relative	.887	.774	.958	.940	.646	.998	.902	1.349	.727
procrustes	.778	.347	.938	.880	.303	.990	.831	1.086	.716
cca	.942	.891	.951	.952	.827	.980	.987	1.224	.895
GW	.802	.053	1.123	.901	.046	1.146	.721	.648	.744
mlp	.932	.680	1.030	.950	.514	1.073	.938	1.360	.748
vec2vec	.829	.000	1.175	.911	.000	1.182	.691	.134	.913
embconv	.897	.625	1.017	.922	.533	1.049	.982	1.074	.971
la2m	.968	.770	1.050	.961	.674	1.095	1.019	1.079	1.003
hmoe	.989	.967	.993	.978	.880	1.022	1.008	1.136	.966
ArguAna									
identity	.591	.000	1.190	.641	.000	1.301	.499	.000	1.000
relative	.925	1.176	.671	.960	1.240	.671	.979	1.001	.957
procrustes	.651	.198	1.112	.685	.159	1.227	.940	.921	.958
cca	.962	.914	1.011	.964	.860	1.070	.996	1.001	.991
GW	.579	.008	1.158	.630	.006	1.272	.556	.167	.945
mlp	.946	.879	1.013	.937	.793	1.085	.984	1.001	.966
vec2vec	.591	.000	1.190	.641	.000	1.301	.499	.000	1.000
embconv	.985	.990	.980	.996	.987	1.005	1.000	1.000	1.000
la2m	.969	.901	1.037	.960	.863	1.059	.999	.999	1.000
hmoe	1.010	1.042	.979	1.015	1.041	.988	.999	.999	1.000
FiQA									
identity	.645	.000	1.101	.722	.000	1.095	.612	.000	1.037
relative	.749	.974	.627	.798	.939	.659	.848	1.093	.683
procrustes	.654	.190	.978	.722	.146	1.005	.780	.758	.787
cca	.956	1.101	.865	.968	1.068	.867	.960	1.094	.874
GW	.631	.023	1.054	.707	.013	1.064	.698	.432	.887
mlp	.976	1.044	.934	.980	.988	.954	.975	1.129	.876
vec2vec	.645	.000	1.101	.722	.000	1.095	.612	.000	1.037
embconv	.980	1.054	.943	.986	1.058	.936	1.000	1.053	.965
la2m	.969	.873	1.039	.957	.820	1.046	1.006	1.011	1.009
hmoe	1.028	1.227	.897	1.035	1.219	.895	.995	1.088	.928

Table A32: Same-family pairs are not easier to translate than cross-family pairs: We group the 24 ordered pairs of Table A3 by how the two encoders are related. (a) Same family, different size: 6 pairs. (b) Shared backbone: 4 pairs, all fine-tuned from Mistral-7B. (c) Different family: 14 pairs. Each cell is a mean full-index `nDCG@10` retention, or the gain over `identity` as a reference, denoted by Δ.

	same family, diff. size		same backbone		different family	
mapper	mean	Δ	mean	Δ	mean	Δ
`identity`	.644	—	.735	—	.653	—
`relative`	.746	.102	.825	.090	.810	.157
`procrustes`	.563	-.081	.633	-.102	.613	-.040
`cca`	.844	.200	.879	.144	.882	.229
`GW`	.571	-.073	.616	-.119	.596	-.057
`mlp`	.832	.188	.870	.135	.894	.241
`vec2vec`	.560	-.083	.629	-.106	.607	-.046
`embconv`	.886	.242	.869	.134	.885	.232
`la2m`	.932	**.288**	.969	**.234**	.937	.284
`hmoe`	.921	.277	.968	.233	.949	**.296**

J RESULTS BY ENCODER RELATIONSHIP

K THE PER-GROUP RESULTS OF TABLE 3

Table 3 averages three groups, each translating two sources into one target index. Table A33 gives every group.

Table A33: Per-group results of Table 3, with the same columns; means over four corpora. (a) and (b) name the source of each D1 half.

	D1				overall	
mapper	(a)	(b)	mixed	D2	ret.	Δ
nv-embed-v2 ← (a) gritlm, (b) gte						
identity	.255	.000	.128	1.121	.694	↓.182
relative	1.027	1.016	1.022	.798	.884	↓.037
procrustes	.144	.182	.163	1.029	.653	↓.175
cca	.763	.843	.803	.967	.887	↓.056
GW	.137	.339	.238	1.116	.743	↓.151
mlp	.777	.828	.803	.995	.913	↓.042
vec2vec	.001	.000	.000	1.121	.636	↓.194
embconv	.600	.642	.621	1.033	.848	↓.087
la2m	.646	.850	.748	1.059	.931	↓.044
hmoe	.818	.933	.876	1.009	.954	↓.029
qwen3-8b ← (a) qwen3-0.6b, (b) gritlm						
identity	.000	.000	.000	1.133	.642	↓.193
relative	.964	.930	.947	.839	.877	↓.042
procrustes	.270	.206	.238	1.057	.699	↓.159
cca	.719	.815	.767	1.003	.898	↓.055
GW	.411	.135	.273	1.095	.739	↓.174
mlp	.672	.744	.708	1.024	.895	↓.056
vec2vec	.000	.000	.000	1.129	.639	↓.193
embconv	.620	.611	.616	1.055	.864	↓.081
la2m	.774	.518	.646	1.076	.895	↓.072
hmoe	.836	.818	.827	1.023	.941	↓.036
gte ← (a) openai, (b) linq						
identity	.000	.000	.000	1.139	.646	↓.193
relative	.910	.617	.764	.859	.824	↓.079
procrustes	.214	.119	.166	1.082	.688	↓.174
cca	.659	.846	.752	1.000	.893	↓.053
GW	.155	.002	.078	1.127	.673	↓.195
mlp	.631	.707	.669	1.020	.882	↓.061
vec2vec	.000	.000	.000	1.135	.643	↓.193
embconv	.659	.636	.648	1.039	.871	↓.075
la2m	.705	.694	.699	1.064	.916	↓.046
hmoe	.795	.910	.853	1.003	.941	↓.032