WHEN ALIGNMENT BECOMES AN INTERFACE: DEPLOYMENT-DEPENDENT INTEROPERABILITY OF TEXT EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Search indexes can outlive the embedding models that created them, leaving documents represented by different encoders. Alignment makes these vectors comparable, but does a method that works well for one document encoder also work when several encoders' documents compete? We study seven frozen text encoders, comparing linear alignment with a shared nonlinear correction across four freshly encoded retrieval datasets and stored vectors for 2.68 million Natural Questions documents. Each document contributes only one encoder's vector. The same correction lowers retrieval quality when all documents use one encoder, yet raises it when documents use a balanced mixture of encoders. On a prospective duplicate-question test, the top-10 relevance score falls by 0.016 in the first setting and rises by 0.103 in the second, on a 0–1 scale. Exchanging score values while preserving each source's document order shows how better competition between sources can outweigh worse ordering within them. The prospective test supports both patterns after adjustment for multiple comparisons. A simpler alternative, standardizing each query's scores against reference texts, improves linear alignment on all four fresh datasets. However, it generally trails nonlinear correction in quality and takes more search time locally. These findings show why alignment should be evaluated on the index it will serve: preserving order within one source and comparing scores across sources are distinct requirements with different quality and cost tradeoffs.

1 INTRODUCTION

Updating a retrieval model need not mean replacing an entire search index. Re-encoding a large document collection takes time and may be impossible when the original text or encoder is unavailable. One option is to align a new query encoder with existing document vectors. Another is to map several encoders into a common space, allowing their vectors to coexist as documents are added or migrated. Work on backward compatibility and embedding conversion has made these options increasingly practical (Cai et al., 2023; Yoon & Arık, 2025; Vejendla, 2025).

These two settings place different demands on alignment (figure 1). When all documents use one encoder, retrieval depends on their order for a given query. When documents use several encoders, the system must also compare scores across those sources. A source whose scores are systematically higher can dominate the merged results even if its own document ordering is unchanged. Thus, success in a single-source index does not determine success in a mixed-source index. We show that this distinction can reverse which alignment method performs better.

Prior research establishes that pretrained representations can be translated, that transfer quality depends on the task and model pair, and that old and new vectors can be searched together (Moschella et al., 2023; Hamideche et al., 2026; Cai et al., 2023). Those results motivate, but do not settle, the comparison studied here: *does the ranking of alignment methods change when the same retrieval task moves from one document encoder to several competing encoders?* We address this question with fixed encoders and relevance judgments, separating changes in within-source order from changes in cross-source score competition. This connects the choice of alignment method to the structure of the index that uses it.

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.

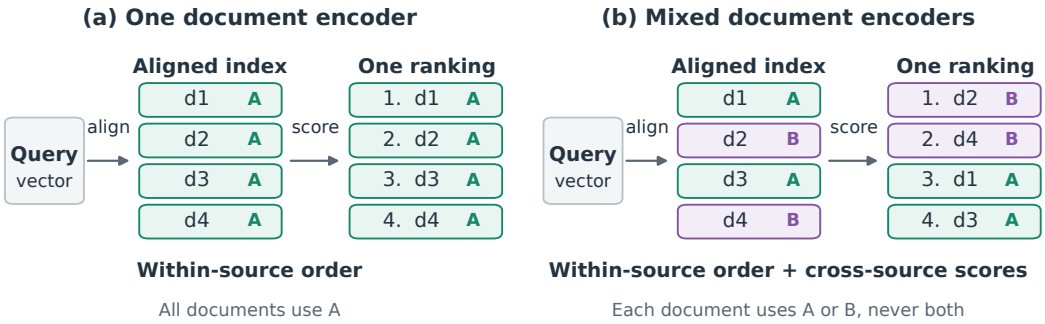


Figure 1: Mixing document encoders adds a score-comparison problem. The same documents use one encoder (left) or one assigned encoder each (right). A/B mark encoder identity, not relevance. After alignment, one query vector scores the index. Mixed retrieval must compare sources as well as order their documents. Rankings are schematic; each document contributes one vector.

Our main finding is a reversal. Across seven text encoders, the nonlinear correction introduced by Achara et al. (2026) makes linear alignment worse for single-source retrieval but better for balanced mixed-source retrieval. The pattern appears on four freshly encoded datasets and a 2.68-million-document Natural Questions collection represented by historical vectors. In every mixed condition, a document contributes exactly one assigned encoder vector; the gains do not come from combining several representations of that document.

The study makes three contributions. First, it establishes the **deployment-dependent reversal**, with matched-weight comparisons, a prospective duplicate-question test, and a large-index check (section 4). Second, an order-preserving score exchange shows a **conditional benefit from cross-source score competition despite a cost to within-source ordering** (section 5). Third, a simple query-dependent score calibration recovers part of the benefit, but its remaining quality gaps and additional searches create a **measurable quality–cost tradeoff** (section 6). Together, these results explain why evaluating alignment on the intended index can lead to a different decision from evaluating encoder compatibility alone.

2 Related Work

Relative representations and latent-space translation enable communication between independently trained models through common anchors or fitted maps (Moschella et al., 2023; Maiorca et al., 2023; 2024). Achara et al. (2026) study multi-way alignment, comparing linear methods with a shared nonlinear correction and examining retrieval, unreliable consensus, incremental model addition, and computational costs. For text encoders, Hamideche et al. (2026) show that translation quality varies across model pairs and downstream tasks. Our study builds on these methods to examine a different source of variation: changing the document-encoder composition of a relevance index while retaining the retrieval task.

Backward-compatible learning trains representations to work with historical features (Shen et al., 2020); L^2R applies compatibility and ranking alignment to continual retrieval with old and newly encoded documents (Cai et al., 2023). Frozen-backbone conversion is also established. Embedding-Converter evaluates both corpus and query conversion, while MKP-Adapter maps new representations into an old gallery (Yoon & Arık, 2025; Byun et al., 2026). Drift-Adapter studies query-to-legacy mappings and migration costs; Efficient Retrieval Adapter (ERA) combines target-corpus alignment with relevance-supervised query adaptation and also evaluates alignment alone (Vejendla, 2025; Maekawa et al., 2026). These studies establish the value of retaining existing vectors. They differ from our comparison in training data, label access, retained indexes, and evaluation targets: for example, Drift-Adapter's text target is the new encoder's exhaustive neighbors rather than human relevance judgments. We compare established aligners under one paired-anchor protocol, rather than claim superiority over these unevaluated converter families. Section K.7 details the closest comparisons.

Score normalization and rank fusion have long addressed the problem of merging results from different retrieval systems (Montague & Aslam, 2001; Cormack et al., 2009). Our mixed index differs from a full-corpus ensemble: each document belongs to one encoder source and receives one score. The question is how alignment changes the competition among those scores. We use established normalization principles as practical controls, and a separate score-exchange diagnostic to distinguish within-source ordering from cross-source competition. The contribution is the observed reversal and its consequences for retrieval quality and serving cost, rather than a new normalization rule or a new mixed-index problem.

3 Retrieval with one or several document encoders

We call an index *homogeneous* when every document uses the same encoder, and *mixed-source* when different documents use different encoders. A source therefore means a document's encoder, not its topic or collection. In either setting, the query encoder can differ from the document encoder. All encoders remain frozen; only the maps between their vector spaces are fitted.

A direct adapter maps a query into an existing document encoder's space, allowing the native index to remain in place. A shared-space alignment instead maps query and document vectors into common coordinates. It can operate on stored vectors without re-encoding document text, but requires corpus projection and usually a new index. These are different operational choices; we use direct adapters as homogeneous references and shared-space methods to compare both index types.

Let $x_i^r(t) \in \mathbb{R}^{1\times d_i}$ be the unit-normalized row vector produced by encoder i for text t, with role $r \in \{q, d\}$ for queries or documents. The role includes the model's prompts, tokenizer, pooling, truncation and precision settings. Write $\mathcal{N}$ for row normalization, with zero vectors remaining zero. The main comparison uses generalized Procrustes alignment (GPA), with $z_i^r(t) = \mathcal{N}(x_i^r(t)W_i)$ and $W_i^\top W_i = I_h$, and geometry-corrected Procrustes alignment (GCPA), which applies a shared nonlinear corrector to GPA features and normalizes the output (Achara et al., 2026). Generalized canonical correlation analysis (GCCA) provides a linear comparison without the orthonormal-map constraint. All three use width $h = 768$. Since this can reduce dimension, GPA need not preserve the original vectors' cosine similarities. Sections A and C.1 give the full definitions and fitting procedures.

Let a be the query encoder and $\alpha(d)$ the encoder assigned to document $d \in \mathcal{D}$. For method m, the mixed score is

$$s^m_{a,\alpha}(q, d) = z_a^{q,m}(q)\, z^{d,m}_{\alpha(d)}(d)^\top. \tag{1}$$

A homogeneous index has constant $\alpha(d) = b$; a mixed index allows α to vary across documents. Each document supplies **exactly one assigned encoder representation**, and each query uses one neural encoder. We cache alternative document encodings to construct experimental assignments, but the scorer never combines them.

Multiplying all scores from one source by a positive constant, or adding an offset, leaves that source's document order unchanged. It can nevertheless change which documents enter the merged top results. This simple distinction separates *within-source ordering* from *cross-source competition*. Neither is a relevance measure by itself. We therefore evaluate the final ranking using normalized discounted cumulative gain at rank 10 (nDCG@10), which rewards judged relevance near the top of the list, and Recall@100, the fraction of judged-positive documents retrieved in the first 100 results. We use linear relevance gains for nDCG and the benchmark convention of zero gain for unjudged documents.

3.1 Experimental design

Encoders and fitting data. We use seven text encoders—EmbeddingGemma, BGE-M3, Arctic, Jina v5, PPLX, Qwen3 and SFR—with native vector widths from 768 to 4096. Each encoder represents the same 10,000 Dolma texts in its document role; these paired *anchors* provide the correspondences for fitting maps. A separate 1,000 anchors supply unlabeled validation, and fits use seeds 42, 43 and 44. Known task-text overlaps are excluded from the final memberships, and no retrieval relevance labels train the maps. Homogeneous references include direct orthogonal and ridge adapters, plus a common frame formed by mapping every encoder into SFR's native space. Detailed recipes and width/anchor controls appear in sections B, C and E.

Retrieval tasks. The initial datasets cover scientific claims (SciFact), finance questions (FiQA), and biomedical search (NFCorpus), using BEIR relevance judgments (Thakur et al., 2021; Wadden et al., 2020; Maia et al., 2018; Boteva et al., 2016). Their document/query counts are 5,183/300, 57,638/648 and 3,633/323, respectively. SciFact and FiQA are retrospective comparisons. NFCorpus was selected and its original comparison registered before scoring; its later diagnostics remain post-hoc. We then test the predicted pattern prospectively on CQADupStack Android, with 22,998 documents and 699 duplicate-question queries (Hoogeveen et al., 2015). Android retains the primary encoder revisions and settings, but exact equality to every earlier tokenizer or remote-code dependency could not be certified; the realized dependencies were pinned before outcomes. Finally, Natural Questions (NQ) supplies a large-index check using stored historical vectors for 2,681,468 documents and 3,452 queries, with a corrector fitted in their compatible historical basis. It is a separate provenance condition, not a fresh re-encoding.

Constructing and comparing indexes. Balanced random assignments distribute documents approximately equally across the seven encoders, independently of relevance labels; all methods use the same assignments. The original homogeneous result averages the 42 directed cross-model pairs and three fits. The original mixed result averages seven query encoders and three assignments using fit seed 42. These summaries differ both in fitting weights and in whether documents can share the query encoder. We report them separately and check matched weights in section 4. Android uses three fits for both index types and three mixed assignments; NQ uses one historical fit and one assignment. Repeated fits and assignments are averaged within each query, rather than treated as additional test queries.

Uncertainty. Paired bootstrap intervals resample the same query groups for compared systems. Original-domain and NQ analyses use 10,000 draws; Android uses 50,000 and groups queries connected by shared positive documents or identical normalized text. All 699 Android groups happen to be singletons. Displayed 95% intervals are *nominal*: they are not adjusted for the number of comparisons. Android's six predeclared directional tests additionally use one-sided Bonferroni bounds at $0.05/6$. We do not pool the different domain summaries or infer equivalence from small differences. Inference is conditional on the fixed corpora, encoders, fitted maps and assignments. Exact scoring accumulates in float64 over float32 vectors and breaks score ties deterministically by corpus row. Sections B and K give the full statistical and exposure protocols.

4 Mixing document encoders reverses the comparison

GCPA's nonlinear correction has opposite effects in the two index settings. It lowers mean relevance quality relative to GPA when the document encoder is fixed, but raises it when documents from different encoders compete (table 1). The result holds across the three original domains. On NFCorpus, for example, the GCPA−GPA difference changes from -0.0277 [-0.0335, -0.0222] in homogeneous retrieval to $+0.0166$ [$+0.0095$, $+0.0238$] in mixed retrieval. Choosing between these methods from homogeneous results alone would therefore select the worse mixed-index method.

Table 1: Correction hurts homogeneous retrieval but helps mixed retrieval. Exact nDCG@10; higher is better. Homogeneous (H) averages 42 directed cross-model pairs and three fits. Mixed (X) averages seven query encoders and three balanced assignments at fit seed 42. Direct adapters retain one native document space. Because the H/X averaging differs, the matched-weight check is reported separately.

Interface	SciFact H	SciFact X	FiQA H	FiQA X	NFCorpus H	NFCorpus X
Direct ridge	0.6430	–	0.3783	–	0.2993	–
Direct orthogonal	0.6800	–	0.4385	–	0.3235	–
GPA	0.6810	0.4613	0.4312	0.2478	0.3210	0.2646
GCCA	0.6872	0.4470	0.4397	0.2464	0.3249	0.2611
GCPA	0.6424	0.5817	0.4183	0.3598	0.2933	0.2812
SFR star	0.6138	0.4256	0.3909	0.2958	0.2786	0.2440

The reversal extends beyond the original domains. Figure 2 adds the prospective Android test and the full historical-vector NQ collection. On Android, the homogeneous difference is -0.0162

[−0.0212, −0.0112], whereas the mixed difference is +0.1027 [+0.0944, +0.1111]. Both directions also pass the adjusted bounds specified before the test. On NQ, the corresponding differences are −0.0501 [−0.0533, −0.0467] and +0.1031 [+0.0984, +0.1079]. The pattern therefore also occurs in a 2.68-million-document index. NQ uses historical vectors and a separately fitted compatible corrector, so it tests the phenomenon in that condition rather than isolating an effect of corpus size.

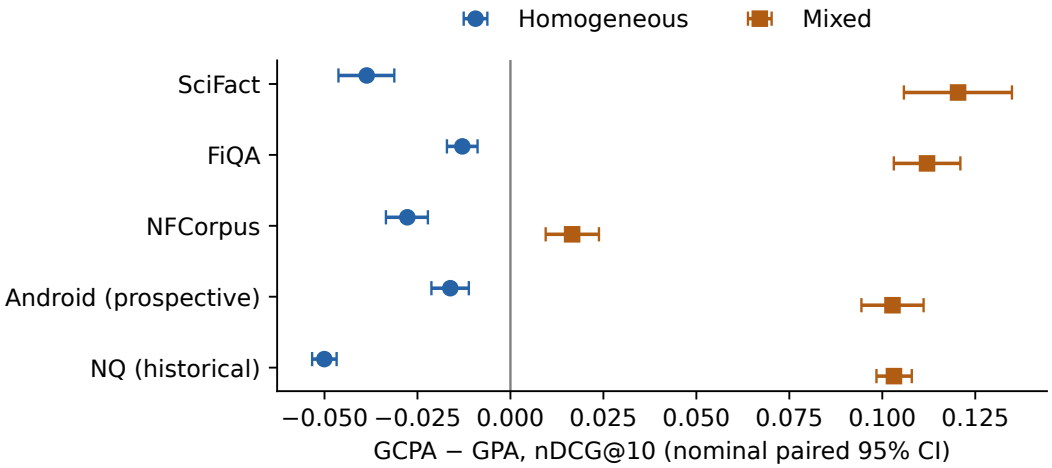


Figure 2: The GCPA−GPA difference reverses sign in all five conditions. Negative values favor GPA; positive values favor GCPA. Bars are nominal paired 95% intervals. The original three domains have different fit weights across deployments; Android uses three fits for both, and historical NQ uses one. Android's adjusted directional tests are reported separately in table 35.

Does the averaging explain the reversal? A post-hoc NFCorpus check uses fit seed 42 for both settings and includes all 49 homogeneous encoder pairs, including same-model pairs. This matches the equal encoder/source weights in the mixed condition. The homogeneous GCPA−GPA difference remains negative ($H_{49} = -0.0301$), while the mixed difference remains positive ($X = +0.0166$). Their difference, $J_{49} = X - H_{49}$, is +0.0467 [+0.0401, +0.0536]. Excluding same-model pairs gives the same conclusion, with $J_{42} = +0.0454$ [+0.0387, +0.0526]. Thus unequal fit weights do not account for this reversal. These are descriptive comparisons of different index constructions; they neither make the candidate representations identical nor replace the original registered summaries (table 23).

What the comparison does and does not recommend. The other aligners reinforce the need to specify the index. Direct orthogonal mapping, GPA and GCCA attain homogeneous NFCorpus means of 0.3235, 0.3210 and 0.3249, yet GCCA's small advantage over GPA does not carry over to mixed retrieval. Averages also hide weak pairs: GPA's FiQA BGE-to-PPLX result is 0.1872 despite a 0.4312 cross-model mean. Full matrices and native/projected same-model references appear in section D.

Nor does the reversal make mixed retrieval preferable to re-encoding every document with a strong model. Native SFR reaches 0.4369 on NFCorpus, compared with mixed GCPA's 0.2812, but requires an SFR vector for every document and SFR query access. The relevant choice here is among methods that can use the available mixed vectors. The next question is why nonlinear correction helps under that constraint while hurting homogeneous retrieval.

5 Separating Document Order from Score Competition

A mixed-index gain can arise in two ways: relevant documents can move up within their source, or sources containing relevant documents can compete more successfully for positions in the merged list. Initial controls suggest that source competition matters. On NFCorpus, GCPA reduces SFR's share of the top 100 results from about 47% to 29% and increases recovery of judged-positive documents from the other six sources. Rankings also change less when encoder assignments are rotated while

document content stays fixed (section F). These observations motivate a more direct test, because a more balanced source mix is not itself evidence of better relevance.

Exchange scores while keeping document order fixed. For each query, we first rank the documents separately within each encoder source, or *shard*. We then keep one method's document order but replace its score values with the other method's sorted scores from the same shard (figure 3). For example, GPA's first-ranked document receives GCPA's largest score, GPA's second-ranked document receives GCPA's second-largest score, and so on. Merging these reassigned scores tests cross-source competition without adopting GCPA's within-source ordering.

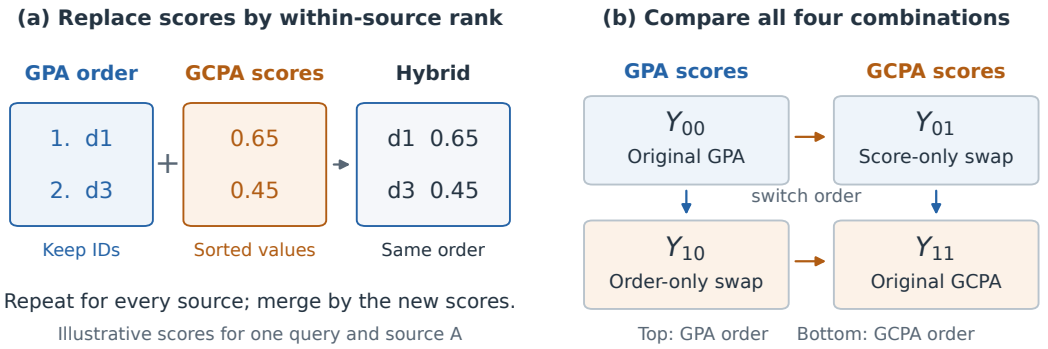


Figure 3: Separating document order from score values. (a) A toy source receives sorted GCPA scores while retaining GPA's document order. Repeat within every source, then merge. (b) Rows choose ordering and columns choose scores; horizontal arrows change scores, vertical arrows change order. The experiment also transfers donor tie-breaking keys to preserve order exactly. Example scores are illustrative, not measured results.

Formally, let $\pi_o(k)$ be the document at rank k under ordering method o, and $v_s(k)$ the kth largest score from donor method s. The *score transplant* assigns

$$\widetilde{s}_{o,s}(q, \pi_o(k)) = v_s(k), \qquad o, s \in \{\text{GPA}, \text{GCPA}\}, \tag{2}$$

then merges the shards. This creates four combinations of order and scores. Write Y_{os} for their relevance quality, using 0 for GPA and 1 for GCPA: Y_{00} and Y_{11} are the original methods, and Y_{01} and Y_{10} are hybrids. The conditional score changes are $S_0 = Y_{01} - Y_{00}$ and $S_1 = Y_{11} - Y_{10}$; the conditional order changes are $O_0 = Y_{10} - Y_{00}$ and $O_1 = Y_{11} - Y_{01}$. Each quantity is first averaged within query over seven encoders and three assignments, at fit seed 42 on the original domains and over three fits on Android.

Preserving order requires care when donated scores tie. We transfer the donor's tie-breaking row key along with its score, preserving the recipient's complete within-shard order and reproducing both original methods exactly. Section J.1 gives the construction and its checks. The diagnostic uses only assigned encoder vectors and no relevance labels, but needs every document's query-specific score. It isolates properties of the resulting rankings; it is neither a cheap serving method nor an intervention on how GCPA was trained.

Better score competition outweighs worse ordering. On NFCorpus, combining GPA order with GCPA scores raises mean nDCG from 0.2646 to 0.2959, exceeding natural GCPA's 0.2812. Conversely, adopting GCPA order lowers quality under either score donor. Table 2 shows the same point-estimate pattern across the other fresh domains, although FiQA's ordering cost under GCPA scores (O_1) is uncertain. On Android, all six predeclared directions pass adjusted bounds: correction hurts homogeneous retrieval, helps mixed retrieval, improves scores under either order, and worsens order under either score donor (table 35). The GPA-order/GCPA-score hybrid is best among the four tested domain-average combinations; this does not make it a theoretical upper bound.

These benefits are conditional and uneven. NFCorpus gains do not extend to Gemma, Jina and PPLX query encoders, and SFR-source recovery declines. The interaction $I = O_1 - O_0 = S_1 - S_0$ also need not be nonzero: its SciFact nDCG interval includes zero, as does its NFCorpus Recall interval. Full contrasts and breakdowns are in section K.2 and table 24. The supported explanation is that a score-value benefit can outweigh an ordering cost under the tested exchange. This suggests a practical

Table 2: GCPA's scores help more than its ordering. Changes in nDCG@10, with nominal paired 95% intervals. S_0 switches GPA to GCPA scores while keeping GPA order; O_0 and O_1 switch to GCPA order under GPA and GCPA scores, respectively. Exact tie handling preserves within-source order. The first three domains are post-hoc; Android is prospective. FiQA's O_1 interval includes zero.

Domain	S_0: scores | GPA order	O_0: order | GPA scores	O_1: order | GCPA scores
SciFact	$+0.1310$ $[+0.1185, +0.1436]$	-0.0129 $[-0.0166, -0.0093]$	-0.0106 $[-0.0150, -0.0065]$
FiQA	$+0.1143$ $[+0.1061, +0.1225]$	-0.0095 $[-0.0115, -0.0076]$	-0.0023 $[-0.0049, +0.0002]$
NFCorpus	$+0.0313$ $[+0.0259, +0.0368]$	-0.0212 $[-0.0252, -0.0175]$	-0.0147 $[-0.0185, -0.0109]$
Android	$+0.1058$ $[+0.0980, +0.1136]$	-0.0076 $[-0.0093, -0.0060]$	-0.0031 $[-0.0053, -0.0008]$

next step: can a simpler calibration improve score competition while leaving GPA's document order intact?

6 Query calibration: quality gains and serving costs

The score transplant uses the full corpus to obtain replacement scores for each query. A practical alternative is to estimate each source's score distribution from a small reference set. We test *query-adaptive calibration* (C1): score the actual query against 512 held document anchors per source, then subtract that source's mean and divide by its standard deviation. This positive affine transformation preserves GPA's within-source order while changing how sources compete. It requires no relevance labels, hyperparameter search or neural fitting.

To distinguish query adaptation from the size of the reference set, a fixed control (C0) uses the same 512-anchor document bank but estimates statistics from 128 document-role query proxies. An earlier fixed rule used a smaller bank and did not reproduce GCPA's benefit. C0 and C1 therefore provide a bank-matched comparison of fixed versus actual-query statistics. Both were studied post-hoc on the original domains; C1 was frozen before the Android test (section J.2).

Table 3: Query-adaptive calibration improves GPA on all four datasets. Exact mixed-source nDCG@10, averaged over seven query encoders and three assignments. C0 uses fixed reference statistics; C1 uses each actual query. Original domains use fit seed 42; Android averages three fits. C0 was not run on Android.

Interface	SciFact	FiQA	NFCorpus	Android
GPA	0.4613	0.2478	0.2646	0.2505
GCPA	0.5817	0.3598	0.2812	0.3532
Fixed (C0)	0.4438	0.2722	0.2600	–
Adaptive (C1)	0.5209	0.3080	0.2806	0.2974

Actual-query statistics recover part of the benefit. C1 improves GPA on all four fresh domains (table 3) and improves C0 on all three where both are evaluated. Its nDCG gains over GPA range from 0.0159 on NFCorpus to 0.0602 on FiQA, with positive nominal paired intervals throughout (table 30 and section K). The improvement shows that nonlinear correction is not the only way to make mixed-source scores more useful.

However, C1 remains below GCPA on SciFact, FiQA and Android. On NFCorpus their nDCG means are close, 0.280558 versus 0.281222, giving C1$-$GCPA $= -0.0007$ $[-0.0076, +0.0062]$ after rounding. Recall@100 is lower for C1 by 0.0166 [0.0106,0.0225]. The close nDCG means therefore do not establish interchangeable performance across metrics, and no equivalence margin was specified. Per-encoder losses and source-recovery tradeoffs remain visible in tables 31 and 32.

Does variance smoothing repair the gap? Dividing by a small estimated standard deviation could make calibration sensitive to the reference bank. We test one prespecified sensitivity: shrink each query/source variance 10% toward the median across sources, holding the means and all other choices

fixed. This lowers nDCG point estimates on all four domains; the SciFact interval includes zero, while the other three are negative (section K.3). The tested smoothing rule does not repair C1's quality gap. It does not, by itself, explain that gap or rule out other calibrators.

Calibration also changes how the index is searched. GPA and GCPA search a single shared-space index. C1's query-dependent source scales and offsets do not fit the unchanged cosine scoring rule of that index. Our implementation instead searches seven source-partitioned GPA indexes and merges the calibrated candidates. Each document still has one vector, but a request now pays for seven searches, reference-bank statistics and merging. An uncalibrated seven-shard GPA baseline separates the cost of partitioning from the additional cost of adaptation. All systems use the same HNSW approximate nearest-neighbor search settings and four total CPU threads.

We time all test queries over five fresh-process passes. The main comparison starts from cached native encoder outputs and includes normalization, projection, GCPA's CPU correction or C1's calibration, all searches and merging. Neural text encoding, networking and queueing are excluded. Table 4 pools raw calls with equal query-encoder/query weights and pairs latency with approximate-search quality under the same document assignment. GCPA is faster on all three measured domains: on FiQA, its median is 1.843 ms versus 8.837 ms for C1. On NFCorpus, C1 has slightly higher approximate-search nDCG, making this a quality–cost tradeoff rather than uniform dominance. The reference bank adds 10.5 MiB of stored document-anchor vectors and requires offline paired-anchor access.

Table 4: C1 takes more search-path time, even when GCPA correction is included. Approximate-search nDCG@10 and warm local latency under assignment seed 42. p50/p95 are the median/95th percentile of raw calls pooled across all seven query encoders. Timing starts from cached native query vectors and includes CPU projection/correction, but excludes neural encoding. Five passes cover every test query.

Domain	GCPA nDCG	GCPA p50/p95 ms	C1 nDCG	C1 p50/p95 ms
SciFact	0.5858	1.100/1.248	0.5276	3.674/3.881
FiQA	0.3655	1.843/2.180	0.3163	8.837/10.085
NFCorpus	0.2797	1.015/1.172	0.2826	3.368/3.605

The practical result is therefore two-sided. Query adaptation improves mixed-index retrieval without changing GPA's document order, but neither its quality nor its cost matches GCPA uniformly. Full timing distributions, the second timing boundary, and bank sensitivity appear in section K. These measurements characterize warm serial local serving; a different indexing or concurrent implementation could change the tradeoff.

7 Discussion and conclusion

The same alignment correction can hurt or help relevance retrieval depending on which encoders supply the documents. This answers the question posed in the introduction: method choice can change when a homogeneous index becomes mixed. The score exchanges explain how the reversal can occur without better within-source ordering, and the calibration experiment shows that some of the gain is recoverable without nonlinear correction. Alignment quality is therefore not a single property that transfers unchanged from one retrieval setting to another.

Prior work makes it possible to reuse representations across model changes; our results show why that reuse should be assessed on the index it will actually serve. An evaluation should measure within-source relevance, cross-source competition, and the operations required to produce the final ranking. It should also distinguish retaining a native index from projecting stored vectors into a new one. Supporting experiments on adding encoders and reusing a fixed classifier further illustrate these distinctions: old maps and vectors can remain unchanged with small observed losses relative to refitting, while classifier transfer still has a substantial gap (section G). Neither geometric agreement nor artifact stability is a substitute for downstream evaluation.

Limitations The study covers seven fixed encoders and mainly balanced random assignments. Real sources may correlate with topic, language or arrival time; testing those allocations is an important next step. Android supplies prospective support under pinned realized dependencies, but

exact equality to every earlier subordinate dependency remains uncertified. NQ supplies a large historical-vector instance, not evidence that corpus size causes the reversal. Known anchor text overlaps were checked, while original document ancestry and encoder pretraining exposure cannot be fully recovered.

The diagnostic exchanges scores after training and does not identify which training component creates the tradeoff. That question requires training-factor ablations. Likewise, C1 tests one practical calibration family, and its measured cost depends on a warm local implementation. Optimized or concurrent serving could alter that cost; measured document throughput also varies with encoder and batch size (section K.6). Whole-corpus migration times remain estimates. These boundaries leave room for better adapters and serving methods without weakening the need to evaluate source competition explicitly.

Conclusion. Alignment should be chosen for the documents that will compete in the final ranking. Across the tested conditions, a correction that loses on homogeneous retrieval wins on mixed retrieval. The diagnostics show how gains in score competition can outweigh losses in document ordering, while query-adaptive calibration offers a simpler, but cost-dependent, alternative. Evaluating these requirements together gives a more useful basis for deploying aligned embeddings than compatibility in a single representation space alone.

REPRODUCIBILITY STATEMENT

Interface definitions, fitting recipes, evaluation roles and statistical estimands are specified in sections A to C and K. The appendices describe preprocessing, calibration, tie handling, uncertainty estimation and timing boundaries. We checked reported intervals and serving percentiles against saved per-query metrics and timings, and retained checksums linking source versions, inputs and failed attempts. Android's subordinate-dependency qualification and the historical-vector status of NQ are explicit; computational reconstruction is distinct from exact encoder reproduction.

AI USE STATEMENT

Generative AI coding assistants were used to implement methods, preprocess data and evaluate code, inspect data and maintain artifact provenance, generate figures, and prepare an initial draft of this manuscript. Numerical results were obtained by executing the recorded experiments rather than by language-model generation; schematic example values are labeled as illustrative. Verification includes independent metric and interval reconstruction, artifact hashes, identity and tie-policy checks, and preservation of failed attempts. The authors are responsible for the final text, scientific interpretation and artifacts; computational checks do not replace that responsibility.

REFERENCES

Akshit Achara, Tatiana Gaintseva, Matéo Mahaut, Pritish Chakraborty, Viktor Stenby Johansson, Melih Barsbey, Emanuele Rodolà, and Donato Crisostomi. Multi-way representation alignment. arXiv:2602.06205v2, 2026. URL https://arxiv.org/abs/2602.06205v2. July 4 revision.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval*, pp. 716–722, 2016. doi: 10.1007/978-3-319-30671-1_58.

Jaeseok Byun, Gukyeong Kwon, Han-Kai Hsu, Meher Gitika Karumuri, Zhikang Zhang, Hao Yang, and Davide Modolo. Multi-modal knowledge preserving adapter for embedding backward compatibility. arXiv:2609.16875v1, 2026. URL https://arxiv.org/abs/2609.16875v1. September 15 version.

Yinqiong Cai, Keping Bi, Yixing Fan, Jiafeng Guo, Wei Chen, and Xueqi Cheng. L2R: Lifelong learning for first-stage retrieval with backward-compatible representations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023. doi: 10.1145/3583780.3614947. URL https://arxiv.org/abs/2308.11512.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, 2020. doi: 10.18653/v1/2020.nlp4convai-1.5. URL https://aclanthology.org/2020.nlp4convai-1.5/.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Sid Ali Hamideche, Louis-Adrien Dufrène, Quentin Lampin, and Guillaume Larue. How far do simple transformations translate across text embedding models? arXiv:2608.05980v2, 2026. URL https://arxiv.org/abs/2608.05980v2. September 2 revision.

Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. CQADupStack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian Document Computing Symposium*, 2015. doi: 10.1145/2838931.2838934.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings. arXiv:2505.12540v4, 2026. URL https://arxiv.org/abs/2505.12540v4.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. arXiv:1702.08734, 2017. URL https://arxiv.org/abs/1702.08734.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Seiji Maekawa, Moin Aminnaseri, Pouya Pezeshkpour, and Estevam Hruschka. Align then adapt: Label-efficient adapter learning for asymmetric dense retrieval. arXiv:2604.03403v2, 2026. URL https://arxiv.org/abs/2604.03403v2.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion of The Web Conference 2018*, pp. 1941–1942, 2018. doi: 10.1145/3184558.3192301.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55394–55414, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ad5fa03c906ca15905144ca3fbf2a768-Abstract-Conference.html.

Valentino Maiorca, Luca Moschella, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via inverse relative projection. arXiv:2406.15057, 2024. URL https://arxiv.org/abs/2406.15057.

Yury A. Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. arXiv:1603.09320v4, 2018. URL https://arxiv.org/abs/1603.09320v4.

Mark Montague and Javed A. Aslam. Relevance score normalization for metasearch. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, 2001. doi: 10.1145/502653.502657.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.15430v2.

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

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/2003.11942.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. arXiv:2402.00159v2, 2024. URL https://arxiv.org/abs/2402.00159v2.

Weiwei Sun, Zhengliang Shi, Jiulong Wu, Lingyong Yan, Xinyu Ma, Yiding Liu, Min Cao, Dawei Yin, and Zhaochun Ren. MAIR: A massive benchmark for evaluating instructed retrieval. arXiv:2410.10127v1, 2024. URL https://arxiv.org/abs/2410.10127v1.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2104.08663.

Harshil Vejendla. Drift-adapter: A practical approach to near zero-downtime embedding model upgrades in vector databases. arXiv:2509.23471v1, 2025. URL https://arxiv.org/abs/2509.23471v1.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Jinsung Yoon and Sercan Ö. Arık. 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, 2025. doi: 10.18653/v1/2025.acl-long.1237. URL https://aclanthology.org/2025.acl-long.1237/.

A Complete interface definitions and cost accounting

A.1 Frozen encoders, preprocessing, and mappings

All feature vectors are **rows**; data matrices stack examples as rows. For encoder $i \in \{1, \dots, M\}$, let $f_i^r : \mathcal{X} \to \mathbb{R}^{1 \times d_i}$ denote the frozen model together with its recipe for role $r \in \{q, d, c\}$ (query, document, classification). A recipe includes tokenizer, prompting/task adapter, truncation, pooling, precision, and model revision. Define $x_i^r(t) = \mathcal{N}(f_i^r(t))$, where $\mathcal{N}(x) = x/\|x\|_2$ for nonzero x and $\mathcal{N}(0) = 0$. The implementation uses a small denominator floor. Role superscripts matter: a map fitted on document-role anchors is not thereby validated for all prompts.

A direct native-target adapter is

$$t_{a\to b}(x) = \mathcal{N}\big((x - \mu_a)T_{a\to b} + \mu_b\big), \qquad T_{a\to b} \in \mathbb{R}^{d_a \times d_b}, \tag{3}$$

with means fitted on normalized training anchors. The corresponding native-target score is

$$s_{a,b}^{\mathrm{dir}}(q, d) = t_{a\to b}(x_a^q(q))\, x_b^d(d)^\top. \tag{4}$$

It transforms queries and can retain a compatible native b index. Availability of the old encoder at query time is unnecessary once the adapter is fitted, but fitting requires paired anchors or access sufficient to create them.

Existing vectors alone do not supply missing correspondences. Unpaired translation, as studied by Jha et al. (2026), addresses a different fitting-access condition; no unpaired result is claimed for this paired-anchor study.

For a shared width h, let $W_i \in \mathbb{R}^{d_i \times h}$. The principal feature maps are

$$\begin{aligned} z_i^{r,\mathrm{GPA}}(t) &= \mathcal{N}(x_i^r(t)W_i^{\mathrm{GPA}}), \\ z_i^{r,\mathrm{GCCA}}(t) &= \mathcal{N}(x_i^r(t)W_i^{\mathrm{GCCA}}), \\ z_i^{r,\mathrm{GCPA}}(t) &= \mathcal{N}\big(C_\theta(x_i^r(t)W_i^{\mathrm{GPA}})\big), \end{aligned} \tag{5}$$

where C_θ is the shared corrector, including its input/output normalization. GPA has column-orthonormal maps for $d_i \geq h$; GCCA allows non-orthogonal maps. The native-star comparator replaces z_i by $t_{i\to b_*}(x_i)$ with a declared native anchor b_* and identity for the anchor itself. A star also requires only $O(M)$ maps; map-count scaling alone does not distinguish it from a learned hub.

A.2 Homogeneous and heterogeneous retrieval

For document set $\mathcal{D}$ and fixed query encoder a, homogeneous hub retrieval uses

$$s_{a,b}^{H,m}(q, d) = z_a^{q,m}(q)\, z_b^{d,m}(d)^\top, \quad d \in \mathcal{D}. \tag{6}$$

The entire index uses document encoder b. For a mixed index, define an assignment $\alpha : \mathcal{D} \to \{1, \dots, M\}$, stored with the documents. The score is

$$s_{a,\alpha}^{X,m}(q, d) = z_a^{q,m}(q)\, z_{\alpha(d)}^{d,m}(d)^\top. \tag{7}$$

Only the assigned vector is supplied to the scorer. The availability of all-model caches to simulate assignments does not authorize using their unassigned rows. A homogeneous index is a constant-assignment special case of this model; its experiment-level aggregate need not equal the mixed experiment's aggregate.

Let $\pi_s(q; \mathcal{D})$ order eligible documents by decreasing score, breaking exact ties by corpus row. Relevance y_{qd} is separate from the encoder identity. We use the campaign's linear-gain discounted cumulative gain,

$$\mathrm{DCG}_{10}(q, \pi) = \sum_{k=1}^{10} \frac{y_{q,\pi_k}}{\log_2(k+1)}, \qquad \mathrm{nDCG@10}(q, \pi) = \frac{\mathrm{DCG}_{10}(q, \pi)}{\mathrm{IDCG}_{10}(q)}, \tag{8}$$

and Recall@100 over documents with positive judgments. Both are averaged over queries. Unjudged documents contribute no gain under the benchmark convention; this does not establish their semantic irrelevance.

Four properties that need not coincide. Representation agreement measures correspondence between two encodings of the same input, for example $z_i(t)z_j(t)^\top$. Ranking preservation concerns whether an interface retains a source's document ordering for a query. Cross-source comparability concerns how those scores compete under α. Relevance retrieval evaluates that combined ordering against judgments. None of the first three definitions contains the relevance judgments that define the fourth.

Two established algebraic facts clarify the distinction. A square orthogonal map preserves dot products. A dimension-reducing map with $W^\top W = I_h$ instead satisfies

$$(xW)(yW)^\top = xWW^\top y^\top. \tag{9}$$

Components outside the retained subspace are lost, and subsequent normalization also matters. Thus our native-to-768 GPA is not a full native-space isometry. Separately, for $a_j > 0$, replacing scores from source j by $a_j s + b_j$ preserves every within-source order but can change the combined ranking. As an illustrative, unmeasured example, scores $0.8, 0.6$ from one source and $0.7, 0.5$ from another interleave differently if the second source receives an offset of 0.2. This possibility motivates calibration controls; it does not prove that an affine discrepancy causes the empirical reversal.

A.3 Fixed classifiers and fixed-hub extension

Let a source-a classifier be $g_a(v) = \arg\max_k (vB_a + b_a)_k$, with its coefficients and intercept frozen after source training/selection. Native-head transfer to encoder b must evaluate

$$g_a\big(t_{b\to a}(x_b^c(t))\big), \tag{10}$$

not apply $T_{a\to b}$ backwards. A head trained on source-a hub features instead receives $z_b^{c,m}(t)$ in the same hub frame. Source-native and target-native fitted heads are distinct reference systems.

For addition, let $U_0 \in \mathbb{R}^{n\times h}$ be the frozen consensus of an old model cohort, with all old maps and indexed vectors fixed. For normalized new-model anchors X_+, GPA-ADD computes

$$W_+ = \operatorname*{arg\,max}_{W^\top W = I_h} \operatorname{tr}(W^\top X_+^\top U_0) = \operatorname{polar}(X_+^\top U_0). \tag{11}$$

The polar factor solves this trace-maximization problem by standard Procrustes algebra (Schönemann, 1966). For dimension reduction, it is not in general the minimizer of $\|X_+W - U_0\|_F^2$: the term $\|X_+W\|_F^2$ can depend on W. This distinction records what the implemented rectangular update actually computes. Holding old representations fixed preserves their exact scores for a fixed query and candidate set. Inserting new candidates can still change ranks, relevance metrics, and the ANN search graph. Artifact invariance is consequently a narrower property than service-output invariance.

A.4 Costs follow access assumptions

For interface m, access state $\mathcal{A}$, and Q future queries, use the additive ledger

$$\begin{aligned} C_m(Q \mid \mathcal{A}) = {} & C_{\text{startup}} + C_{\text{anchor encoding}} + C_{\text{fit}} + C_{\text{doc encoding}} \\ & + C_{\text{projection}} + C_{\text{index}} + Q\big(c_{\text{query encoding}} + c_{\text{mapping}} + c_{\text{search}}\big). \end{aligned} \tag{12}$$

Terms are charged only when required by $\mathcal{A}$. A direct adapter may reuse an existing native index. A newly adopted hub normally projects stored vectors and builds a new index, but needs no corpus text. Re-encoding into a replacement native model requires corpus text and encoder access even if cached replacement vectors supply an offline quality reference. ADD to an existing hub fits one map and processes new documents; it does not pay for the old hub again. Separate-index retrieval uses one neural query encoder but several maps/searches and a merge. Memory, index bytes, and latency percentiles are recorded alongside this ledger; time components are not interchangeable with hardware cost or monetary charges.

B Full fitting, exposure, and statistical protocol

Models and data. We use EmbeddingGemma, BGE-M3, Snowflake Arctic, Jina v5, PPLX-Embed, Qwen3-Embedding, and SFR-Embedding, spanning native widths 768–4096. Model identifiers,

revision prefixes and encoding settings are reported in section C. Canonical SciFact, FiQA, and NFCorpus arrays were freshly encoded with pinned primary model recipes and stored as float32. This canonical condition is distinct from the historical-vector NQ scale check. SciFact (Wadden et al., 2020) contains 5,183 documents and 300 test queries; FiQA (Maia et al., 2018), 57,638 and 648; NFCorpus (Boteva et al., 2016), 3,633 and 323. We use the BEIR relevance-retrieval formulation (Thakur et al., 2021).

SciFact and FiQA are retrospective source domains: earlier project outcomes were exposed. NFCorpus was selected before its preparation/scoring as an external project-outcome check, with official development/test roles and frozen artifacts before test execution. The audit checked that all 50 external jobs began after registration. This establishes chronology for the original comparisons, not absence from encoder pretraining. NFCorpus is now exposed and cannot freshly confirm a mechanism proposed after reviewing these results.

Fitting. The main maps use 10,000 paired document-role Dolma anchors and 1,000 separate unlabeled validation anchors, with fitting seeds 42, 43, and 44. Known task-text matches and available duplicate text groups are excluded by the r3 memberships. The underlying overlap audit identified 1,800 matching windows in the historical 100,000-row pool. Generated window IDs and missing original subset tags prevent a guarantee of disjoint original documents or entities. No evaluation relevance labels train the primary maps.

Direct ridge and rectangular orthogonal adapters operate at the actual target width, retaining fitted means. Ridge minimizes $n^{-1}\|X_aT - X_b\|_F^2 + \lambda\|T\|_F^2$ on centered normalized anchors, choosing $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ by unlabeled validation correspondence error. Orthogonal fitting uses the rectangular polar factor; widening maps cannot have orthonormal columns. Native-star maps use the same ridge family into SFR, a declared anchor rather than a test-selected optimum. Gemma/Qwen stars are source-domain sensitivity controls.

GPA/GCCA/GCPA use width 768. We reference the pinned implementation of Achara et al. (2026), with explicit differences in native-to-hub width handling, input row normalization, and disjoint corrector validation anchors. The normalized-consensus GPA updates and exact GCPA training recipe are given in section C.1; we do not assert that this dimension-reducing implementation inherits every guarantee of square GPA. Upstream architecture, loss, gradients, and a 40-update optimizer replay were checked, the latter with maximum parameter discrepancy zero on the tested partition. Exact quality scoring uses float64 accumulation over float32 representations with deterministic tie handling; ANN uses float32. The 38 empty FiQA documents remain eligible, including recorded zero encoder outputs for empty text.

Primary aggregates. For method m, let $L^m_{q,a,b,s}$ be per-query homogeneous nDCG for pair $a \neq b$ and fitting seed s. Define

$$\widehat{\theta}^m_H = \frac{1}{|\mathcal{Q}|}\sum_q \frac{1}{3M(M-1)} \sum_{s\in\{42,43,44\}} \sum_{a\neq b} L^m_{q,a,b,s}, \tag{13}$$

$$\widehat{\theta}^m_X = \frac{1}{|\mathcal{Q}|}\sum_q \frac{1}{3M} \sum_{r\in\{42,43,44\}} \sum_{a} L^m_{q,a,\alpha_r,42}. \tag{14}$$

Balanced assignments α_r distribute corpus rows approximately equally among seven sources, independently of labels, and are shared across methods. The mixed aggregate includes query encoders matching the source of some documents; the homogeneous aggregate excludes diagonals. Mixed fitting-seed checks and nested Gemma-to-SFR migration are supporting conditions, not replacements for equation (14).

Headline contrasts average replicas within each query, then use 10,000 paired bootstrap resamples of normalized-query-text groups. *Paired* means the same query groups are sampled for the compared systems. *Nominal 95%* means the intervals are not adjusted for the number of comparisons and do not provide 95% simultaneous coverage across all reported contrasts. Intervals are conditional on the fixed corpus, model cohort, fitted maps, and finite assignments. Forty-two directed pairs are not 42 independent model samples, and repeated fitting seeds are not additional test queries. Classifier intervals analogously resample example groups; macro-F1 is recomputed from resampled predictions.

We specify no equivalence, noninferiority, or acceptable-loss margin. The ADD aggregate has a different design: seven held-out cohorts, twelve incident directions each, at fitting seed 42.

Separately frozen post-hoc follow-up. After inspecting the original results, we froze an additional NFCorpus-only protocol for a four-arm score-order transplant and matched-weight deployment contrasts. It uses fitting seed 42, all seven query encoders, the three original balanced assignments, and the unchanged test queries/corpus. Its 84 cells and 21 identity conditions belong to a separate source/output snapshot. It trains no maps, loads no neural encoders, and acquires no new data. The two identity arms must reproduce all saved top-100 rows before interpreting any transplant. This gate and a subsequent independent metric audit passed. The follow-up's nominal paired intervals use the same 10,000 group draws across all primary, secondary, and matched contrasts. Its exposure and protocol status remain post-hoc throughout.

A subsequent, separately frozen review-revision protocol tests exact tie-policy sensitivity (84 additional cells) and two calibration rules across the three already exposed domains (126 new quality cells, with 189 original reference cells). It also measures 105 cached-feature ANN conditions at assignment 42. Both phases retain all prespecified conditions and require independent reconstruction; no maps or neural encoders are fitted in that revision. These analyses do not extend the original external registration. The later Android, source-diagnostic, serving and historical-NQ extensions have their own protocols in section K; architectural/loss ablations remain unexecuted.

Computational verification. The completed core audit covers 2,583 homogeneous cells, 1,274 mixed cells, 434 addition-related cells, and 714 classifier cells, with separately identified width/anchor controls and 23 full-NQ cases. Follow-up/revision cells are additional, not replacements for the original campaign. These counts include references and supporting conditions; they are not inferential sample sizes. During manuscript preparation, we linked exhibits to source hashes and reproduced six original reversal intervals, 30 first-follow-up intervals, and 124 review-revision intervals from saved per-query metrics. Audit consistency is evidence that the declared computation ran, not proof of semantic validity or causal identification.

C Recipes and reproducibility

Table 5: Canonical encoder configuration, including primary model revision prefixes. Caps are configured upper bounds, additionally limited by the loaded model's sequence limit. Model identifiers and query/document settings are described below.

Encoder	Width	Precision	Token cap	Model revision prefix
Gemma	768	bf16	2048	`57c266a740f5`
BGE	1024	fp16	8192	`5617a9f61b02`
Arctic	1024	fp16	8192	`ac6544c8a46e`
Jina	1024	fp16	32768	`d153fb80cb0b`
PPLX	2560	bf16	32768	`06456497a005`
Qwen	4096	bf16	32768	`1d8ad4ca9b3d`
SFR	4096	bf16	4096	`f62d15f411ca`

The exact repositories are `google/embeddinggemma-300m`, `BAAI/bge-m3`, `Snowflake/snowflake-arctic-embed-l-v2.0`, `jinaai/jina-embeddings-v5-text-small`, `perplexity-ai/pplx-embed-v1-4b`, `Qwen/Qwen3-Embedding-8B`, and `Salesforce/SFR-Embedding-2_R`. Table prefixes are display abbreviations only; artifact identities use full hashes.

Gemma uses its document/query prompt names. Arctic uses the query prompt name. Jina uses the retrieval task with the query prompt on queries. BGE and PPLX use their recorded default recipes. Qwen and SFR use dataset-specific instruction templates, including the biomedical NFCorpus instruction frozen in the external registration. The whitespace after `Query:` is model-specific and retained. Stored canonical outputs are float32 even where encoder inference uses reduced precision. Empty documents are eligible: native normalization of a zero vector remains zero, whereas a declared affine/nonlinear map may produce a nonzero vector from a fitted bias.

We retained versioned numerical snapshots, model configurations and checksums linking each exhibit to its source records. The original completion audit byte-verified 14,158 main objects (21,040,789,375 bytes), with separate full-index and source-code checks. These records support computational traceability, not statistical guarantees. The procedures, settings and principal verification results are described in this appendix.

C.1 Fitting details and standard algebra

The paired anchor pool consists of Dolma windows (Soldaini et al., 2024). Row memberships are separately recorded for each fit; overlapping windows across fitting replicas are not independent datasets. The r3 exclusions change memberships, not previously encoded values. Train and validation roles are disjoint under available text-group evidence, with the ancestry limitations in the main text.

For normalized anchor matrices X_i and consensus U, the canonical GPA loop updates

$$W_i \leftarrow \mathrm{polar}(X_i^\top U), \qquad Z_i \leftarrow \mathcal{N}(X_i W_i), \qquad U \leftarrow \mathcal{N}\left(M^{-1}\sum_i Z_i\right), \tag{15}$$

where normalization applies rowwise. It runs at most 50 iterations and stops when the mean cosine objective changes by less than 10^{-5}. Float64 SVD computes polar factors, followed by float32 storage. These normalized-consensus iterations describe the realized GPA variant, rather than claiming globally optimal minimization of a rectangular least-squares objective. GCCA calls the pinned upstream shared-rank alignment routine, pads only for that routine's handling of unequal widths, and retains the native rows of its returned maps. Inference applies the fitted operators and normalization, with no silently substituted centering or whitening.

GCPA uses the shared upstream MLP at hidden multiplier 2 and dropout 0. For normalized GPA feature u and normalized paired consensus c, its sampled loss is

$$\ell(u, c) = 1 - C_\theta(u)c^\top + 0.7\max\{0, 1 - C_\theta(u)u^\top - 0.08\}. \tag{16}$$

Training uses AdamW at learning rate 0.002 and weight decay 10^{-4}, gradient clipping at 1, batch size 4,096, and a maximum of 50×200 sampled updates. Validation is checked every 20 updates using 50 sampled validation batches; 50 non-improving checks trigger stopping, and the best state is restored. The corrector is applied to both query and document paths. The parity checks cover architecture/state transfer, loss/gradients, and a fixed 40-update optimizer partition, not the validity of an entire semantic model of alignment.

For a rectangular cross-matrix $K = L\Sigma V^\top$, the implemented polar factor is $LV^\top$. The trace-maximization statement in equation (11) follows from the singular-value trace inequality; non-uniqueness at rank deficiency does not require inventing missing singular directions beyond the solver's completion. Under compression, column orthonormality does not imply native dot-product preservation or an invertible native round trip. We use these established facts solely to interpret the implementation.

C.2 Exact affine calibration and source-local merging

For a fixed query encoder, abbreviate the document-source calibration constants as $a_b = 1/\sigma_b > 0$ and $b_b = -\mu_b/\sigma_b$. Given row query z_q and document z_d assigned to b, define augmented rows

$$\widetilde{z}_q = [z_q, 1], \qquad \widetilde{z}_d = [a_b z_d, b_b]. \tag{17}$$

Their inner product is $a_b z_q z_d^\top + b_b$. These augmented rows must *not* be unit-normalized again, because that changes the intended score. For a separate-index merge, a document outside its source's top k has at least k same-source documents ranked ahead under a positive affine transform. It therefore cannot enter global top k, subject to consistent tie ordering. This is why local top-100 collection gives exact calibrated top-100 merging in the control.

RRF assigns each document a sum of $1/(60 + r)$ over returned ranks r. Here shards are disjoint, so a document receives one term, and ties across shards use global corpus row. This encourages interleaving similarly ranked documents from each populated source. It does not estimate their calibrated relevance probabilities and is sensitive to how the corpus is partitioned. All unassigned alternative document embeddings remain inaccessible to these scorers.

C.3 Uncertainty, exposure, and remaining inferential gaps

Within a normalized-query group g, let S_g be the sum of query-level paired contrasts and n_g its query count. Each bootstrap draw samples groups with replacement and computes $\sum_g w_g S_g / \sum_g w_g n_g$. Thus duplicate-text groups move together while the estimand remains query-weighted. Percentiles use 10,000 draws and seed 42. These intervals do not include a random-effects model over encoders or fitted maps.

The ADD contrast averages all seven held-out cohorts and both incident directions before this bootstrap. A directed pair appears in two different held-out-hub contexts; these are valid distinct conditions, not independent observations. Classifier macro-F1 is recomputed under example-group resampling for each fixed fitting replica and then averaged, retaining all 77 classes. Accuracy uses the mean of correctness indicators.

The external registration freezes model/method choices, fits, allocation seeds, and calibration/ANN settings before the original NFCorpus outcomes. No later reinterpretation extends that registration. The completed follow-up in section I supplies an explicitly post-hoc score-order diagnostic and matched-weight method-by-deployment contrasts. A causal decomposition of GCPA training remains unidentified. Original and follow-up artifacts have separate protocol, source, audit, and completion identities.

D Homogeneous retrieval detail

Table 6: Original homogeneous nDCG@10 means, 42 off-diagonal pairs and three fitting seeds. Native-target direct maps and 768-dimensional hub methods have different corpus-update requirements.

Interface	SciFact	FiQA	NFCorpus
Direct ridge	0.6430	0.3783	0.2993
Direct orthogonal	0.6800	0.4385	0.3235
GPA	0.6810	0.4312	0.3210
GCCA	0.6872	0.4397	0.3249
GCPA	0.6424	0.4183	0.2933
SFR star	0.6138	0.3909	0.2786

Table 7: Off-diagonal homogeneous means, mean Recall@100, and minimum seed-averaged pair. Each row includes the same 42 directed pairs and three fits; "minimum" is not a confidence bound.

Domain	Interface	nDCG	Recall	Minimum	Worst pair
SciFact	Direct ridge	0.6430	0.9118	0.2338	BGE→SFR
SciFact	Direct orthogonal	0.6800	0.9348	0.4889	BGE→SFR
SciFact	GPA	0.6810	0.9362	0.5422	BGE→SFR
SciFact	GCCA	0.6872	0.9374	0.5508	BGE→SFR
SciFact	GCPA	0.6424	0.9228	0.5080	BGE→Gemma
SciFact	SFR star	0.6138	0.8948	0.2338	BGE→SFR
FiQA	Direct ridge	0.3783	0.7196	0.0487	BGE→SFR
FiQA	Direct orthogonal	0.4385	0.7717	0.3145	BGE→PPLX
FiQA	GPA	0.4312	0.7650	0.1872	BGE→PPLX
FiQA	GCCA	0.4397	0.7725	0.2039	BGE→PPLX
FiQA	GCPA	0.4183	0.7561	0.2966	BGE→Gemma
FiQA	SFR star	0.3909	0.7278	0.0487	BGE→SFR
NFCorpus	Direct ridge	0.2993	0.3069	0.1405	BGE→SFR
NFCorpus	Direct orthogonal	0.3235	0.3210	0.2061	BGE→SFR
NFCorpus	GPA	0.3210	0.3202	0.2188	BGE→PPLX
NFCorpus	GCCA	0.3249	0.3238	0.2186	BGE→PPLX
NFCorpus	GCPA	0.2933	0.3056	0.2071	BGE→SFR
NFCorpus	SFR star	0.2786	0.2960	0.1405	BGE→SFR

Table 8: Original three-domain differences underlying the first three rows of figure 2. Nominal paired intervals for the separately defined homogeneous and mixed averages; the matched-weight comparison is in table 23.

Domain	Homogeneous Δ [95% CI]	Mixed Δ [95% CI]
SciFact	-0.0387 $[-0.0462, -0.0312]$	$+0.1204$ $[+0.1058, +0.1349]$
FiQA	-0.0129 $[-0.0171, -0.0088]$	$+0.1120$ $[+0.1031, +0.1210]$
NFCorpus	-0.0277 $[-0.0335, -0.0222]$	$+0.0166$ $[+0.0095, +0.0238]$

Table 9: Native and within-model projected nDCG@10, averaged over the three fitting seeds. These diagonal references are excluded from the homogeneous cross-model mean.

Domain	Encoder	Native	GPA	GCCA	GCPA
SciFact	Gemma	0.7868	0.7868	0.7835	0.7432
SciFact	BGE	0.6436	0.6416	0.6397	0.5991
SciFact	Arctic	0.7098	0.7099	0.7057	0.6812
SciFact	Jina	0.7647	0.7655	0.7638	0.7288
SciFact	PPLX	0.7581	0.7539	0.7523	0.7315
SciFact	Qwen	0.7852	0.7823	0.7796	0.7508
SciFact	SFR	0.8593	0.8555	0.8472	0.8167
FiQA	Gemma	0.4739	0.4739	0.4719	0.4490
FiQA	BGE	0.4128	0.4124	0.4102	0.3725
FiQA	Arctic	0.4536	0.4540	0.4530	0.4192
FiQA	Jina	0.4944	0.4927	0.4896	0.4740
FiQA	PPLX	0.5749	0.5743	0.5742	0.5572
FiQA	Qwen	0.6471	0.6394	0.6388	0.6104
FiQA	SFR	0.6203	0.6117	0.6098	0.5812
NFCorpus	Gemma	0.3933	0.3929	0.3885	0.3541
NFCorpus	BGE	0.3146	0.3152	0.3075	0.2775
NFCorpus	Arctic	0.3528	0.3521	0.3513	0.3133
NFCorpus	Jina	0.3978	0.3963	0.3975	0.3677
NFCorpus	PPLX	0.3820	0.3765	0.3726	0.3441
NFCorpus	Qwen	0.4194	0.4138	0.4143	0.3769
NFCorpus	SFR	0.4369	0.4261	0.4258	0.3905

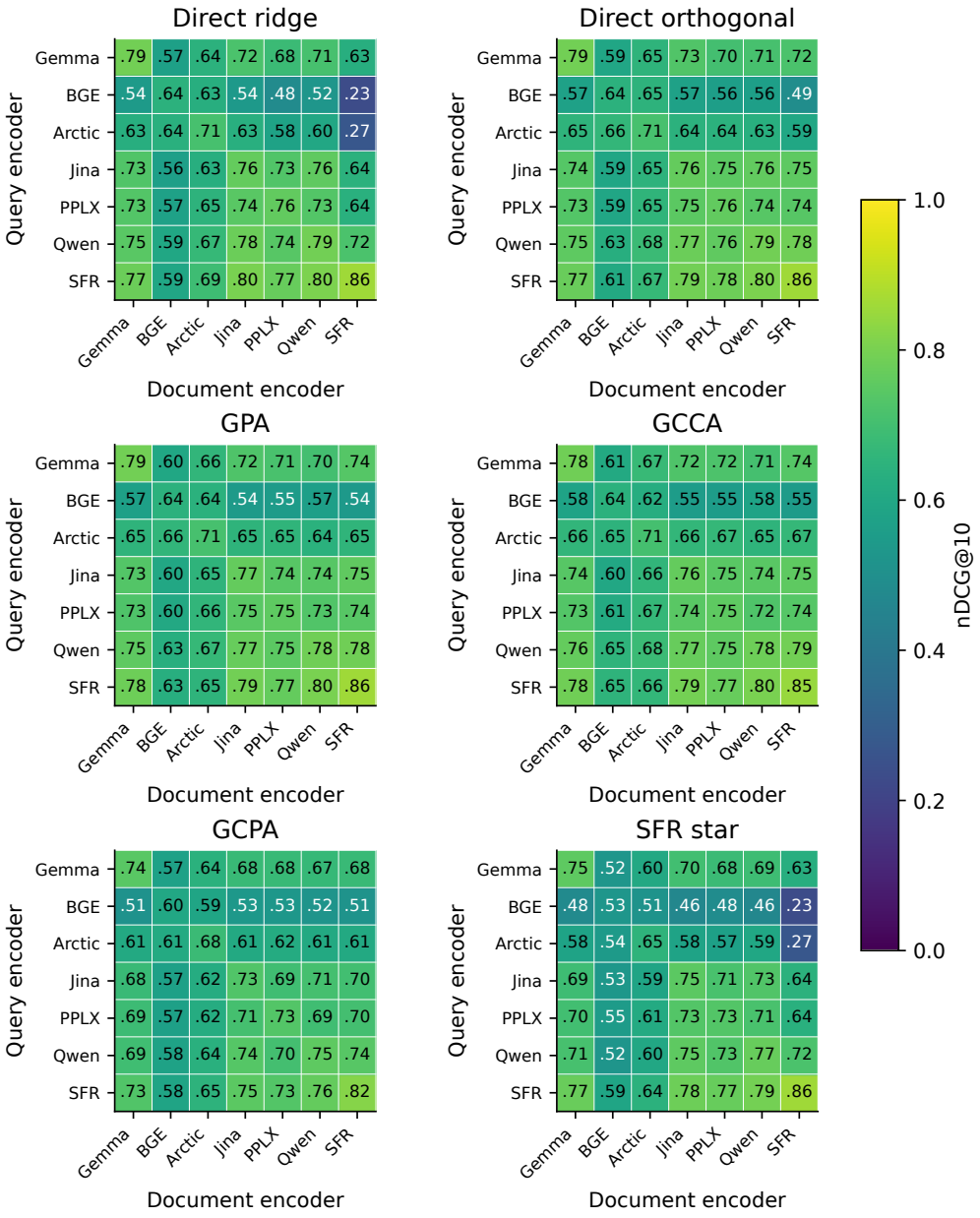


Figure 4: SciFact compatibility matrices, three-fit means. Direct-map diagonals display native references; hub/star diagonals display projected same-model scores. Shared color scale $[0, 1]$. Cell annotations are rounded to two decimals for readability.

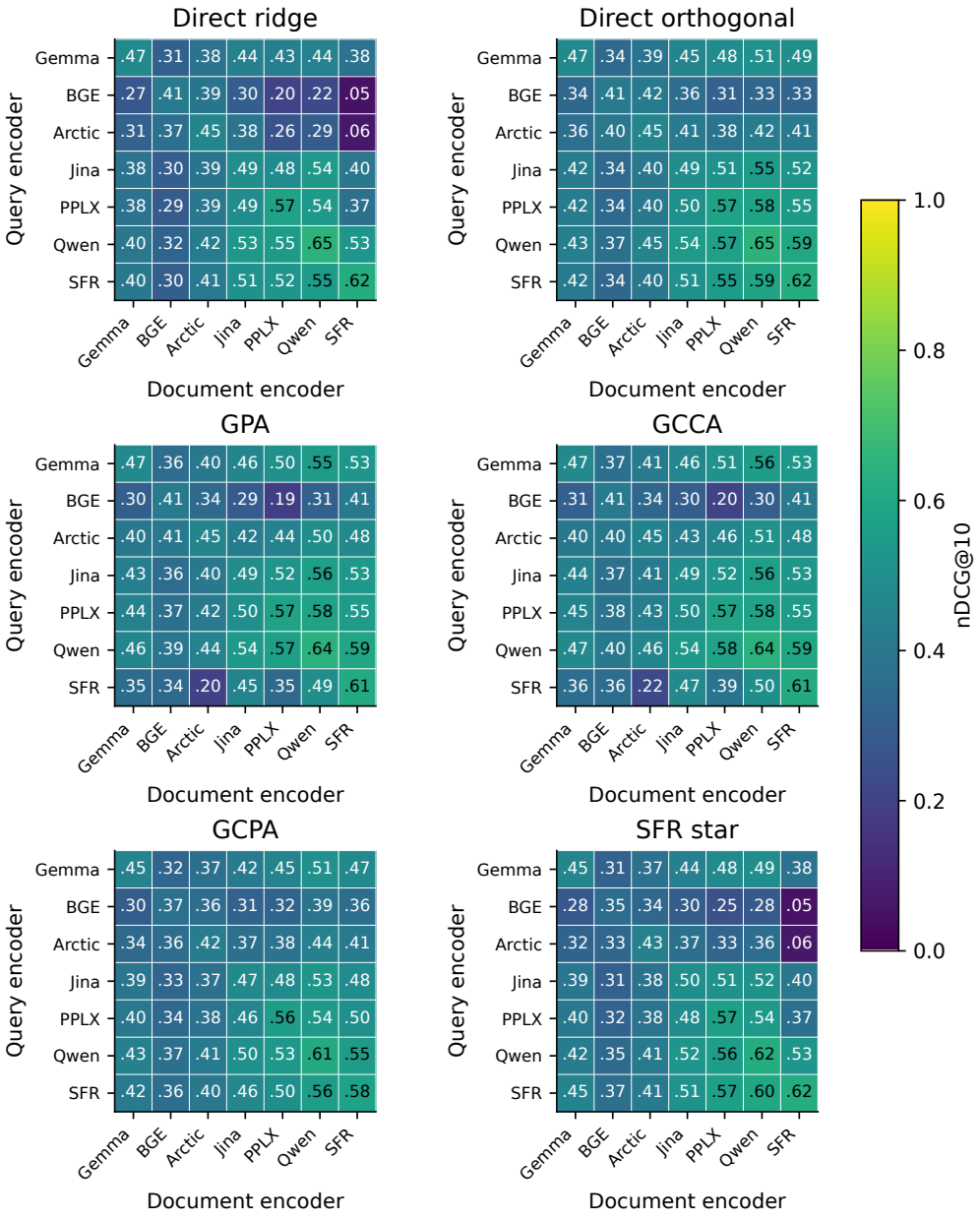


Figure 5: FiQA compatibility matrices with the same conventions. The low BGE-to-PPLX GPA cell illustrates why a cohort mean does not guarantee pair-level usefulness.

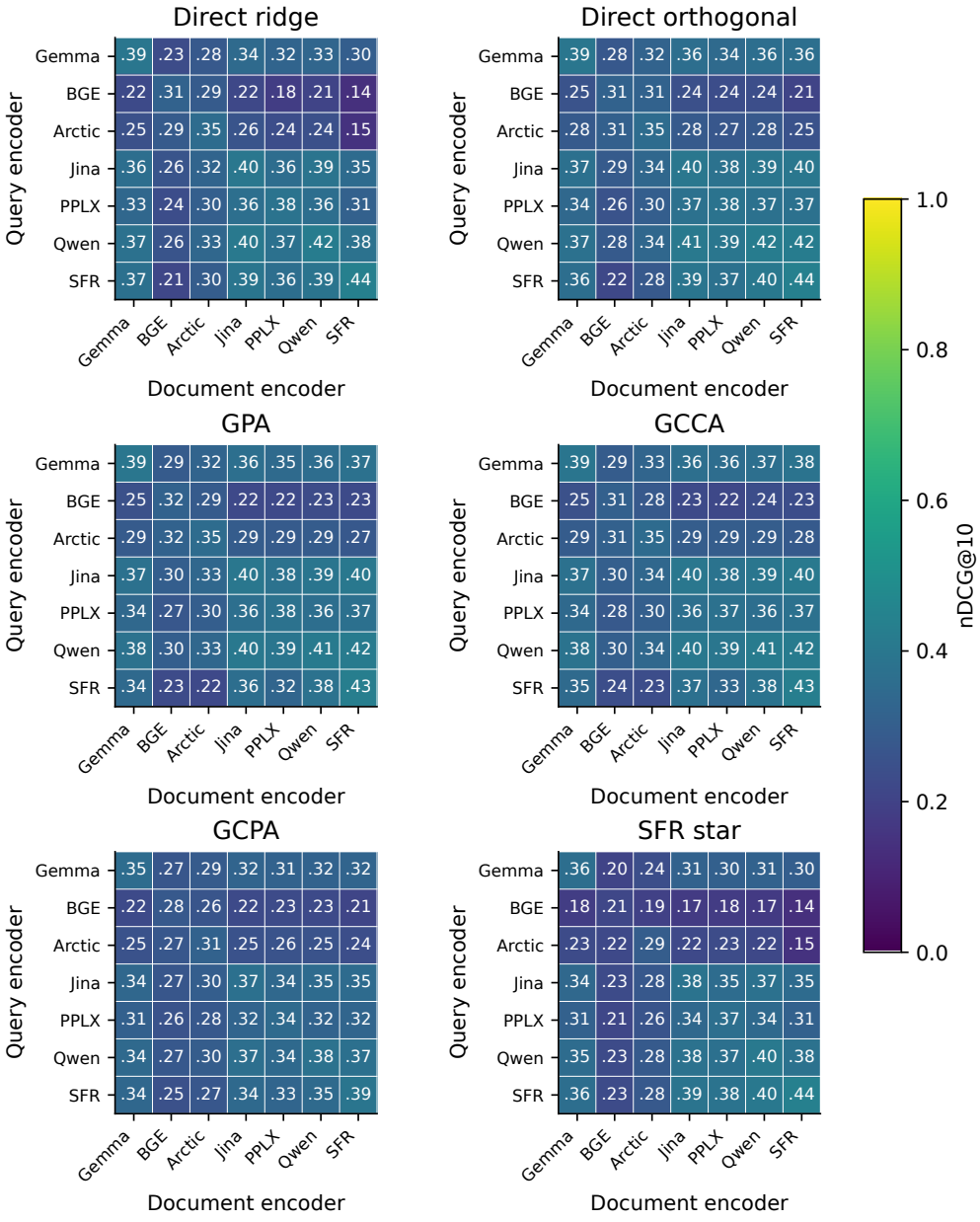


Figure 6: NFCorpus compatibility matrices for the registered external check, with the same conventions.

E MIGRATION AND MATCHED-WIDTH CONTROLS

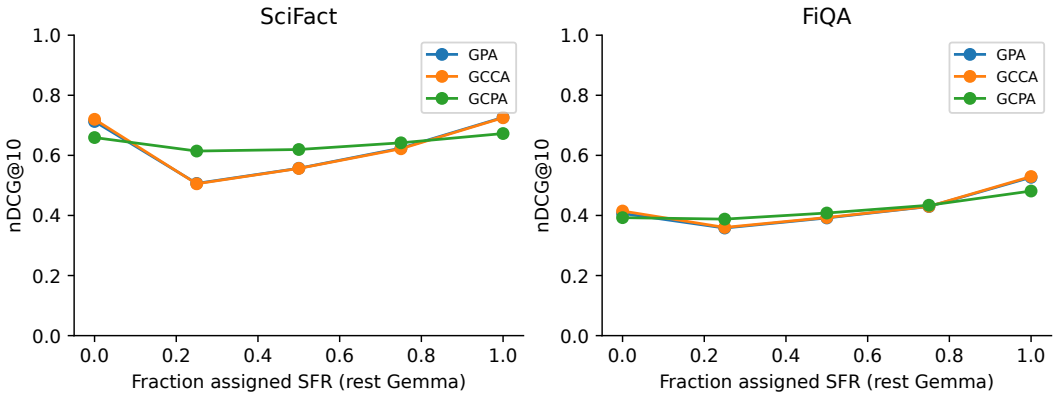


Figure 7: Nested partial migration, fitting/assignment seed 42. Each point averages seven query encoders. A fixed random ordering chooses the SFR prefix; every other document uses Gemma. Lines connect measured fractions only. These two-source assignments differ from balanced seven-source retrieval.

Direct native-target systems and hub768 systems differ in both mapping and output width. The saved controls fit PCA768 and native adapters on the same paired-anchor protocol, and also examine GPA widths 256/512 and a 1,000-anchor condition. PCA projection changes the corpus and requires indexing the transformed vectors; it is not unchanged-native-index reuse. Controls are retrospective and not a full width-by-budget sweep. The generated matched off-diagonal table avoids mixing the seven native diagonals with the 42 transfer cells.

Table 10: Supporting width/anchor controls, fitting seed 42, 42 off-diagonal cells per row. Native-target and PCA-target baselines have different corpus-update requirements.

Condition	SciFact	FiQA
native-ridge	0.6422	0.3763
native-orthogonal	0.6800	0.4380
gpa-d768-n10000	0.6805	0.4304
pca-ridge	0.6606	0.4326
pca-orthogonal	0.6734	0.4533
gpa-d256-n10000	0.6321	0.3821
gpa-d512-n10000	0.6715	0.4220
gpa-d768-n1000	0.5999	0.3476

F ADDITIONAL MIXED-INDEX CONTROLS

Three controls help interpret the mixed-index reversal: fixed score normalization, separate-index merging, and reassignment of encoders to unchanged documents. They test whether simple score adjustments recover quality and whether the retrieved list depends strongly on which encoder represents each document. These original experiments motivated the later score-exchange and query-adaptive tests; none identifies a unique training mechanism.

F.1 FIXED AFFINE CALIBRATION AND SEPARATE INDEXES

Table 11: Original balanced mixed-source means: seven query encoders and three assignments, fitting seed42. Each document contributes one vector. The fixed affine control uses held corpus-role anchor statistics; it differs from the later C0/C1 bank protocol.

Interface	SciFact	FiQA	NFCorpus
GPA	0.4613	0.2478	0.2646
GCCA	0.4470	0.2464	0.2611
GCPA	0.5817	0.3598	0.2812
SFR star	0.4256	0.2958	0.2440
Separate RRF	0.3965	0.2574	0.2195
Separate calibrated	0.4045	0.2444	0.2355
GPA calibrated	0.4482	0.2747	0.2617

For each query-encoder/document-source pair, the original affine control estimates mean/std from all cross-products of 128 held corpus-role anchors on the query side and a disjoint 128 on the document side, without relevance labels. Augmented inner-product scoring or positive-affine source-local merging reproduces the score exactly (section C.2). Calibration improves GPA on FiQA by 0.0269 [0.0204,0.0337] but does not clearly improve SciFact or NFCorpus. The NFCorpus change is -0.0029 $[-0.0079, +0.0021]$. This fixed rule is insufficient in those runs; it does not rule out the later actual-query calibrator.

Separate RRF reaches 0.2195 on NFCorpus, below GPA's 0.2646 and GCPA's 0.2812. Documents occur in only one native shard, so each receives one rank contribution, unlike full-corpus ensemble fusion. Nearly equal source shares do not demonstrate better relevance merging. Separate calibrated retrieval reaches 0.2355; its improvement over RRF still leaves it below the common-frame alternatives here. These systems use one neural query encoder but several native-target maps/searches.

F.2 SOURCE RECOVERY AND REASSIGNMENT

Table 12: Original controls. Overlap is mean intersection/100 between balanced assignment42 and a cyclic label rotation with identical document content/IDs. The final column is the three-assignment calibrated-GPA minus GPA nDCG contrast with nominal paired95% CI.

Domain	GPA overlap	GCPA overlap	Calibrated GPA − GPA
SciFact	0.3153	0.5049	-0.0130 $[-0.0298, +0.0034]$
FiQA	0.2190	0.4199	$+0.0269$ $[+0.0204, +0.0337]$
NFCorpus	0.2889	0.4755	-0.0029 $[-0.0079, +0.0021]$

On NFCorpus, GPA retrieves SFR-source documents with mean top100 share 47.0% under an approximately 14.3% corpus share. SFR is also the best native encoder; the descriptive seven-model Spearman correlation between native quality and GPA source share is 0.3214. These observations neither prove harmful bias nor explain all source preference.

GCPA lowers SFR share to 28.8% and its positive-pair recovery from 0.3572 to 0.2812, while improving recovery for the other six document sources. This is a relevance tradeoff, not a Pareto gain or cosmetic balancing success. Background-score samples include unjudged documents and are not treated as known negatives.

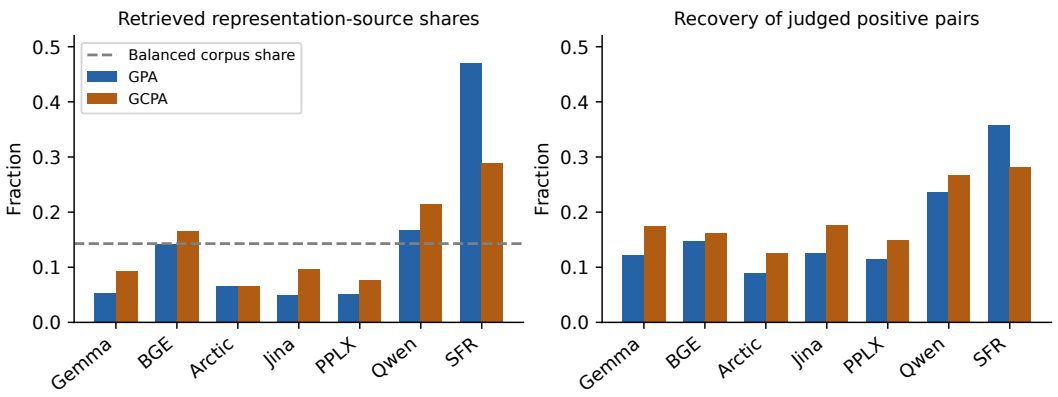


Figure 8: NFCorpus, fitting/assignment seed42, averaged over seven query encoders. Left: top100 source shares. Right: micro recovery of judged-positive query–document pairs within source, then averaged over query encoders. This is not query-macro Recall@100 or the later three-assignment profile.

The label-rotation control changes which encoder represents each fixed document. NFCorpus top100 overlap increases from 0.2889 for GPA to 0.4755 for GCPA; the same ordering holds on SciFact and FiQA. Approximately half the GCPA list still changes. Rank displacement is calculated only for documents present in both lists; absent documents are censored. These results are consistent with reduced representation-source dependence under correction, while leaving several explanations open. Nested Gemma-to-SFR migration in figure 7 similarly shows that intermediate allocations need not interpolate homogeneous endpoints. It is a five-point supporting experiment, not an optimized migration policy or an observed temporal deployment.

G Extending and reusing a fixed interface

G.1 New encoders without moving old vectors

An existing shared index is easier to extend if adding an encoder leaves its old maps and vectors intact. We test the quality cost of this constraint using the incremental alignment of Achara et al. (2026). For each of the seven encoders, we fit a six-model GPA hub without it, freeze that hub, and fit only the missing encoder's map to the existing consensus (ADD). The comparison is a full seven-model refit (REFIT). Evaluation covers all twelve new-to-old and old-to-new directions for each held-out encoder. A native-star reference uses SFR as its fixed common space, or Qwen when SFR is held out; this choice is specified in advance.

The differences are small in absolute nDCG units (table 13): ADD reaches 0.6802 versus 0.6805 for REFIT on SciFact, and 0.4294 versus 0.4304 on FiQA. SciFact's ADD−REFIT interval includes zero; FiQA's is negative. Without a prespecified tolerance, these results support "small observed difference," not equivalence or acceptable loss. They are source-domain results, with no NFCorpus ADD confirmation.

Old maps and old corpus vectors were byte-identical under ADD. The separate insertion experiment also verified the old indexed vectors after new candidates were inserted. These checks support stability of stored representations. They do not imply an unchanged ANN graph or unchanged results after the candidate set grows. The insertion results are evaluated as a different retrieval problem.

Median new-map fitting takes 0.061–0.082 seconds across encoders on the measured H100 environment, with serialized maps of 2.25–12.00 MiB (table 16). The timing includes processing/transferring cached new-model anchors, but excludes neural anchor generation, old-hub construction, and new-document indexing. Its operational implication is that the incremental algebra can be inexpensive *when the shared interface and paired anchors already exist*. It is not a claim that encoder onboarding costs a fraction of a second end to end.

Table 13: Fixed-hub addition: nDCG@10 over seven held-out cohorts and twelve incident directions each, fitting seed 42. Paired query-group intervals average cohort/direction contrasts within query. REFIT is the matched full-seven GPA reference. Native star follows the conditional anchor rule.

Domain	ADD	REFIT	Native star	Δ ADD$-$REFIT [95% CI]
SciFact	0.6802	0.6805	0.6253	-0.0003 $[-0.0011, +0.0006]$
FiQA	0.4294	0.4304	0.4021	-0.0010 $[-0.0016, -0.0004]$

Table 14: Banking77 fixed-source head reuse. Cross results average 42 directions and three map-fitting seeds. Same-model results average seven encoders and three seeds; direct adapters use the native head (native same-model accuracy 0.8872). A dash avoids inventing a direct same-model fit. Corpus-role map fitting and classification-role encoding make this an explicit cross-role condition.

Interface	Cross accuracy	Cross macro-F1	Same accuracy
Direct ridge	0.6403	0.6225	–
Direct orthogonal	0.6676	0.6519	–
GPA	0.6575	0.6402	0.8872
GCCA	0.6708	0.6542	0.8870
GCPA	0.6701	0.6530	0.8829

G.2 Classifier reuse as corroborating evidence

A second reuse question concerns a classifier trained on one encoder's features: can it accept another encoder's aligned features without retraining? Banking77 (Casanueva et al., 2020) supplies a limited test of whether this task favors the same alignment changes as homogeneous retrieval. For each source encoder and feature system, we fit a logistic classifier, select regularization using source features, then freeze its coefficients and intercept. Other encoders are evaluated through the correctly directed map in equation (10). Each classifier uses 616 training labels (eight per class), 2,000 selection labels and 3,080 evaluation examples. Separately trained target-native classifiers use the same labeled IDs but require target-side fitting.

GCPA improves mean cross-model accuracy over GPA by 0.0126 [0.0103, 0.0150], reaching 0.6701 versus 0.6575 (table 14). GCCA is 0.6708 and direct orthogonal 0.6676; these close point estimates do not establish equivalence. All are well below approximately 0.887 same-model performance. Per-pair signs vary substantially (section G.3), so the average is not a guarantee for any particular replacement.

This is evidence that fixed-head utility and homogeneous relevance retrieval reward different changes in the interface, consistent with the deployment focus. It is not evidence that GCPA has solved classifier migration. Banking77 is one fixed 77-class taxonomy, the source label budget includes selection, and the mappings were trained on corpus-role rather than classification-role anchors. These distinctions preclude a general prompt-invariance or new-taxonomy transfer claim.

Table 15: Addition by held-out encoder, twelve incident directions per row, fitting seed 42. Different rows use different six-model hubs. The conditional native-star anchor changes to Qwen only when SFR is held out.

Domain	Held out	ADD	REFIT	Native star
SciFact	Gemma	0.6929	0.6933	0.6432
SciFact	BGE	0.5920	0.5928	0.4925
SciFact	Arctic	0.6500	0.6511	0.5590
SciFact	Jina	0.7031	0.7030	0.6561
SciFact	PPLX	0.6963	0.6966	0.6564
SciFact	Qwen	0.7088	0.7086	0.6679
SciFact	SFR	0.7182	0.7178	0.7020
FiQA	Gemma	0.4279	0.4303	0.3926
FiQA	BGE	0.3362	0.3384	0.2881
FiQA	Arctic	0.4023	0.4039	0.3379
FiQA	Jina	0.4544	0.4549	0.4242
FiQA	PPLX	0.4499	0.4500	0.4273
FiQA	Qwen	0.4986	0.4989	0.4632
FiQA	SFR	0.4362	0.4360	0.4814

Table 16: Three-repeat medians for fitting only the new map from cached anchors to a frozen consensus, including host-to-GPU transfer. Neural anchor encoding, old-hub construction, and new-document processing are excluded. Serialized map bytes include the array header.

Added encoder	New map fit (s)	Map (MiB)
Gemma	0.0783	2.25
BGE	0.0608	3.00
Arctic	0.0605	3.00
Jina	0.0605	3.00
PPLX	0.0693	7.50
Qwen	0.0822	12.00
SFR	0.0817	12.00

G.3 FIXED-SOURCE CLASSIFIER DIRECTIONS

Table 17: Banking77, source head to target encoder, three-fit averages. The feature-map direction for a native source head is target-to-source. GCPA−GPA intervals are nominal paired example-group intervals. Pair-level signs differ; no multiple-comparison-adjusted discovery claim is made.

Source	Target	GPA	GCPA	Δ accuracy [95% CI]
Gemma	BGE	0.6677	0.7561	$+0.0883$ [$+0.0766$, $+0.1004$]
Gemma	Arctic	0.7569	0.7524	-0.0045 [-0.0121, $+0.0029$]
Gemma	Jina	0.3134	0.2971	-0.0163 [-0.0244, -0.0082]
Gemma	PPLX	0.8130	0.8076	-0.0054 [-0.0120, $+0.0012$]
Gemma	Qwen	0.8458	0.8367	-0.0091 [-0.0158, -0.0026]
Gemma	SFR	0.6729	0.7746	$+0.1016$ [$+0.0891$, $+0.1141$]
BGE	Gemma	0.8485	0.8425	-0.0060 [-0.0117, -0.0001]
BGE	Arctic	0.8159	0.8123	-0.0036 [-0.0098, $+0.0028$]
BGE	Jina	0.2652	0.2227	-0.0424 [-0.0506, -0.0341]
BGE	PPLX	0.7878	0.7936	$+0.0058$ [-0.0016, $+0.0134$]
BGE	Qwen	0.8002	0.7899	-0.0103 [-0.0167, -0.0042]
BGE	SFR	0.6727	0.7153	$+0.0425$ [$+0.0341$, $+0.0512$]
Arctic	Gemma	0.7790	0.8302	$+0.0512$ [$+0.0425$, $+0.0598$]
Arctic	BGE	0.6947	0.7999	$+0.1052$ [$+0.0932$, $+0.1172$]
Arctic	Jina	0.3235	0.2984	-0.0251 [-0.0351, -0.0152]
Arctic	PPLX	0.7124	0.7562	$+0.0437$ [$+0.0355$, $+0.0519$]
Arctic	Qwen	0.7421	0.8018	$+0.0597$ [$+0.0500$, $+0.0697$]
Arctic	SFR	0.4174	0.7202	$+0.3028$ [$+0.2862$, $+0.3190$]
Jina	Gemma	0.6218	0.5631	-0.0587 [-0.0690, -0.0485]
Jina	BGE	0.3091	0.3907	$+0.0816$ [$+0.0690$, $+0.0942$]
Jina	Arctic	0.3947	0.3380	-0.0567 [-0.0666, -0.0469]
Jina	PPLX	0.5588	0.4657	-0.0931 [-0.1038, -0.0827]
Jina	Qwen	0.6103	0.5220	-0.0883 [-0.0992, -0.0775]
Jina	SFR	0.2737	0.3773	$+0.1036$ [$+0.0905$, $+0.1166$]
PPLX	Gemma	0.8434	0.8539	$+0.0105$ [$+0.0047$, $+0.0165$]
PPLX	BGE	0.6284	0.7419	$+0.1135$ [$+0.1018$, $+0.1257$]
PPLX	Arctic	0.7383	0.7424	$+0.0041$ [-0.0038, $+0.0121$]
PPLX	Jina	0.2192	0.1732	-0.0460 [-0.0534, -0.0386]
PPLX	Qwen	0.8338	0.8285	-0.0053 [-0.0116, $+0.0008$]
PPLX	SFR	0.6263	0.7752	$+0.1489$ [$+0.1364$, $+0.1615$]
Qwen	Gemma	0.8881	0.8814	-0.0067 [-0.0118, -0.0017]
Qwen	BGE	0.7259	0.7762	$+0.0503$ [$+0.0406$, $+0.0601$]
Qwen	Arctic	0.7785	0.7780	-0.0004 [-0.0083, $+0.0074$]
Qwen	Jina	0.3852	0.2828	-0.1024 [-0.1126, -0.0927]
Qwen	PPLX	0.8504	0.8380	-0.0124 [-0.0183, -0.0068]
Qwen	SFR	0.8105	0.8365	$+0.0260$ [$+0.0183$, $+0.0337$]
SFR	Gemma	0.8857	0.8775	-0.0082 [-0.0134, -0.0030]
SFR	BGE	0.7181	0.7262	$+0.0081$ [-0.0004, $+0.0166$]
SFR	Arctic	0.7945	0.7608	-0.0337 [-0.0419, -0.0254]
SFR	Jina	0.4788	0.3319	-0.1469 [-0.1596, -0.1341]
SFR	PPLX	0.8472	0.8208	-0.0264 [-0.0332, -0.0198]
SFR	Qwen	0.8648	0.8552	-0.0096 [-0.0153, -0.0041]

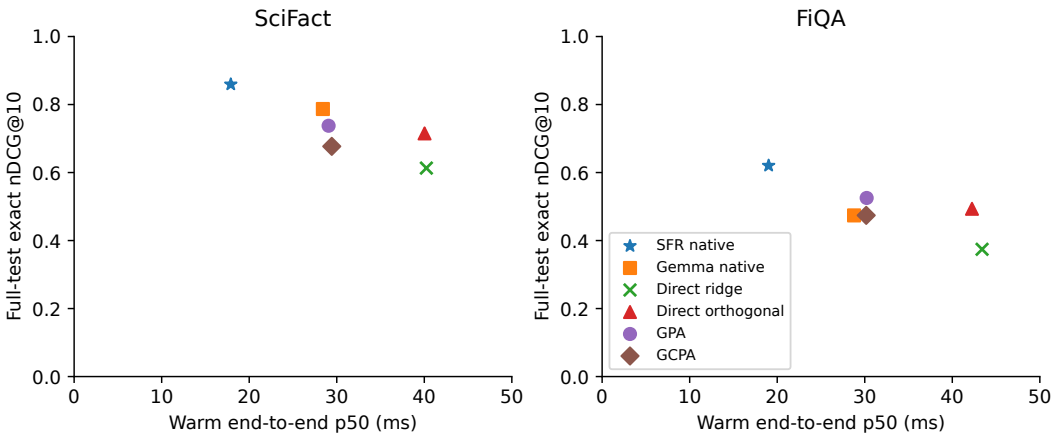


Figure 9: Declared Gemma-query/SFR-document deployment example with both native references. Quality is full-test exact nDCG at fitting seed 42; latency is measured warm development-query end-to-end p50. Direct ridge/orthogonal use the SFR native index; GPA/GCPA use projected SFR vectors in a new 768-dimensional index. Plotting these quantities together is a deployment comparison, not a joint confidence region.

H Original operational profiles and historical-vector scale

H.1 A small map is not necessarily a fast service

This earlier service profile compares the operations required to reuse a native index with those required to adopt a shared-space index. Unlike the expanded calibration benchmark, it includes neural query encoding, but uses a small development-query panel. We profile all seven encoders using one H100, CPU-resident maps/corrector, and FAISS CPU HNSW (Johnson et al., 2017; Malkov & Yashunin, 2018). Each online measurement uses batch size one, two warmup calls, and three repetitions of the first sixteen accepted development queries: 48 timed calls per system/domain. Mapping and search use four CPU threads. HNSW uses $M = 32$, `efConstruction`=100 and `efSearch`=400; development curves also include 100/200. Startup is timed separately. These are warm local measurements rather than production service-level guarantees.

The predeclared small-to-large example is Gemma queries over SFR document vectors. On FiQA, direct orthogonal mapping gives 0.4935 exact nDCG and 42.25/44.61 ms p50/p95; GPA gives 0.5252 and 30.22/31.06 ms, with index size 183.82 rather than 915.56 MiB. GPA is attractive for this measured combination of quality, latency, and index size *if* corpus projection/rebuilding is permitted. Direct mapping retains a distinct advantage under an immutable-native-index requirement. Their feasible operations differ before any Pareto comparison is made.

Parameter count is a poor latency proxy in this harness. Native SFR's FiQA p50 is 19.02 ms, versus 28.78 ms for native Gemma. The larger model is faster on this workload, so we do not describe Gemma as a measured faster query encoder. Direct native-width mapping also incurs nontrivial CPU overhead; charging only its parameter count would conceal that overhead. These observations are conditional on recipes, lengths, implementation, and hardware, not a general architecture ranking. Full p50/p95 and index values appear in table 18.

H.2 Migration ledgers rather than free cached replacements

Equation (12) separates work that is otherwise easy to conflate. Reuse of a native index avoids its rebuild; possession only of vectors may still require an index build. Hub adoption can avoid neural document encoding while requiring projection and a new index. Cached a/a vectors provide a replacement-quality reference but do not remove the encoding cost of a real replacement. Similarly, a parameter-free vector operation is not free access to a neural encoder.

The recorded ledgers report measured setup components and estimates at 10^4, 10^6, and 10^8 future queries (table 19 shows one example). Query costs use the measured *mean*, not the median; setup divided by query volume is added to that mean. Neural document encoding is extrapolated from eight

character-length strata, up to eight documents per batch and three timing repeats. It is not a measured whole-corpus migration or optimized throughput forecast. Paired-anchor generation is outside the cached-anchor ledger, and some direct-family fit timings conservatively include unused target fits. These limitations prevent interpreting the ledger as a deployment-independent break-even guarantee.

H.3 Adaptive calibration changes the serving path

Query-dependent source offsets/scales cannot be treated as ordinary cosine scores in the same single 768-dimensional index. We implement C1 with seven source-partitioned GPA HNSW indexes, using one assigned vector per document, and charge calibration, all searches, and merging. We compare it with single-index GPA/GCPA and an uncalibrated seven-shard GPA merge; C0 shares those shard indexes. This distinguishes the cost of partitioning from the additional query-adaptive statistics.

The original calibration benchmark uses assignment 42, four CPU threads, HNSW $M = 32$, efConstruction100 and efSearch400, with 48 timed calls per query-encoder/domain/system after two warmups. All systems receive cached *final* query features, uniformly excluding neural encoding and query projection/correction. The combined pipeline is timed directly. On NFCorpus, the median across seven encoder-specific p50s is 2.976 ms for C1-shards, 1.067 ms for GPA-shards, 0.388 ms for GPA-single, and 0.409 ms for GCPA-single. On FiQA these medians are 7.717, 5.330, 0.936 and 1.019 ms. These historical summaries are not pooled percentiles (table 33); the larger-panel correction-inclusive study is in section K.4.

C1's source document-bank payload is 10.5 MiB, with setup and index costs recorded separately. All stored vectors in the 27 persisted indexes were byte-checked. ANN relevance and top100 agreement are compared with each system's own exact calibrated ranking on the full test set. Their assignment42 means differ from the three-assignment exact averages in table 3. The evidence is consequently a domain/metric/cost tradeoff: a simple adaptive rule substantially improves GPA, but seven searches and calibration cost more in this implementation. Lower cached-feature GCPA search time does not include the cost of producing its corrected features.

H.4 Full NQ: approximation error is distinct from alignment loss

A small-corpus result need not describe index behavior at scale. We therefore evaluate 23 full-NQ cases, using the retrieval adaptation of Natural Questions (Kwiatkowski et al., 2019), on 2,681,468 documents and 3,452 queries: six prescribed directed pairs under ridge, orthogonal mapping, and GPA; four native references; and one balanced seven-source GPA allocation with Gemma queries. NQ is a **historical-frozen-vector condition**. Its paired anchors were cleaned and refitted at the primary width/budget, but legacy vector metadata do not bind current encoder revisions. It validates operations on those exact historical arrays, not fresh reproducibility by today's encoder releases.

For Gemma-to-SFR, GPA's exact nDCG is 0.5778 versus 0.5187 for direct orthogonal and 0.2826 for ridge. Native SFR is 0.7211 and native Gemma 0.6347. At `efSearch`=400, GPA attains 0.5749 with 0.9728 top-100 recall against exact and 1.52 ms search-only p50. The original mixed GPA case reaches 0.3412 exact and 0.3383 ANN400. This original check omitted GCPA; the completed matched homogeneous/mixed GCPA extension is reported separately in section K.5.

The original six pairs were Gemma↔SFR, BGE↔Qwen and Qwen↔SFR, chosen to cover small/large and large/large interfaces rather than random, exhaustive or worst-case pair sampling. The executed scale protocol preserves this purposeful cohort. An immutable timestamp proving that this choice preceded every historical NQ outcome was not recovered, so we do not recast it as a fresh external registration.

The 768-dimensional NQ indexes occupy 8.35 GiB versus 41.60 GiB for 4096-dimensional native indexes, including HNSW overhead. Native-4096 builds reached approximately 98.44 GiB process RSS versus approximately 19.62 GiB for the 768-dimensional builds, so index bytes alone understate construction memory. ANN fidelity also depends on the interface: all-case curves in figure 10 and the complete results in table 20 report agreement with each method's own exact ranking. Lower ANN loss does not repair an already poor aligned exact ranking. The supplemental audit checked every stored vector in all nine full-NQ indexes against its declared projection, supporting this separation between representation choice and approximate search. The hardware/runtime ledger records submitted resources, nodes, images, and process measurements. Source-domain timings used H100 encoding

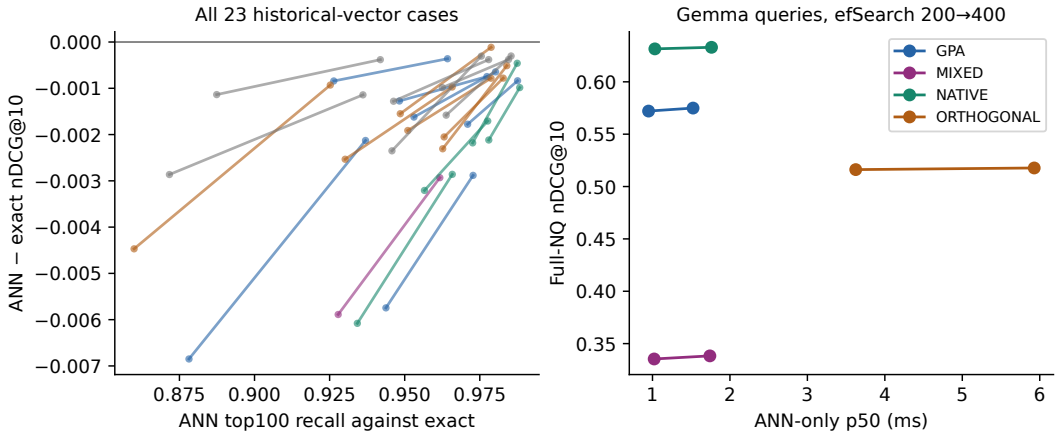


Figure 10: Full historical-vector NQ. Each segment connects the two measured `efSearch` settings, 200 and 400, for one fixed interface. Left: all 23 cases, separating ANN top-100 agreement from downstream nDCG loss. Right: Gemma queries for selected native, SFR-target, and mixed systems. Search timings use cached query vectors and exclude neural encoding; only two settings are measured.

Table 18: Warm service example. Latency: first 16 development queries, three repeats, batch size one, CPU maps and HNSW400. Quality: full test exact scoring, seed42; it is not the ANN metric measured on the timing panel. SFR native is the retained-old-service reference when that encoder remains available. Gemma native is a replacement-quality reference with a different corpus representation.

Domain	System	Test nDCG	p50 / p95 (ms)	Index (MiB)
SciFact	SFR native	0.8593	17.89 / 22.17	82.33
SciFact	Gemma native	0.7868	28.42 / 30.87	16.53
SciFact	Direct ridge	0.6131	40.24 / 41.04	82.33
SciFact	Direct orthogonal	0.7151	40.03 / 41.34	82.33
SciFact	GPA	0.7376	29.07 / 29.49	16.53
SciFact	GCPA	0.6770	29.43 / 29.62	16.53
FiQA	SFR native	0.6203	19.02 / 23.93	915.56
FiQA	Gemma native	0.4739	28.78 / 30.72	183.82
FiQA	Direct ridge	0.3743	43.40 / 45.14	915.56
FiQA	Direct orthogonal	0.4935	42.25 / 44.61	915.56
FiQA	GPA	0.5252	30.22 / 31.06	183.82
FiQA	GCPA	0.4740	30.17 / 30.66	183.82

and CPU search; the large native HNSW builds used CPU-only allocations of 16 CPUs and 192 GiB. Full-NQ evaluation used four FAISS threads. NQ p50/p95 derive from 20 cached single-query searches after two warmup calls, whereas ANN quality uses all 3,452 queries. Build-stage wall time includes integrity checks and I/O and is not an isolated arithmetic kernel measurement. The full-NQ exact scorer streams bounded corpus blocks; simultaneous resident raw embeddings of all seven models are not required.

Table 19: FiQA/Gemma-query ledger conditional on an existing SFR native index and cached paired anchors. Setup includes startup and measured fit costs in addition to the shown corpus costs; encoding is extrapolated. Direct methods reuse the index. The final column is estimated amortized time at one million serial queries, using measured mean online time, not p50.

System	Setup (s)	Encode est. (s)	Project (s)	Build (s)	At 10^6 (ms/q)
Gemma replacement	302.42	284.84	0.14	6.85	29.81
ridge	23.87	0.00	0.00	0.00	43.40
orthogonal	23.87	0.00	0.00	0.00	42.71
gpa	46.31	0.00	1.31	6.32	30.35
gcpa	62.40	0.00	2.40	5.60	30.24

Table 20: All 23 full-NQ historical-frozen-vector cases. Exact/ANN400 columns are nDCG@10; ANN recall is top-100 agreement with that method's exact ranking, not relevance Recall@100. Latencies are search-only, using 20 timed cached queries; quality uses the full 3,452-query set.

Method	Query→document	Exact	ANN400	ANN recall	Search p50/p95 ms
GPA	BGE→Qwen	0.3692	0.3671	0.9370	1.83 / 2.63
GPA	Gemma→SFR	0.5778	0.5749	0.9728	1.52 / 2.22
GPA	Qwen→BGE	0.5456	0.5448	0.9774	1.84 / 2.51
GPA	Qwen→SFR	0.6532	0.6524	0.9876	1.56 / 2.40
GPA	SFR→Gemma	0.5564	0.5561	0.9643	1.79 / 2.77
GPA	SFR→Qwen	0.5208	0.5202	0.9802	1.82 / 2.73
mixed	Gemma→mixed7	0.3412	0.3383	0.9617	1.74 / 2.35
native	BGE→BGE	0.6061	0.6032	0.9658	2.52 / 3.05
native	Gemma→Gemma	0.6347	0.6330	0.9776	1.76 / 2.68
native	Qwen→Qwen	0.6464	0.6455	0.9883	5.60 / 8.14
native	SFR→SFR	0.7211	0.7206	0.9875	6.20 / 8.23
orthogonal	BGE→Qwen	0.3648	0.3639	0.9252	6.96 / 9.72
orthogonal	Gemma→SFR	0.5187	0.5177	0.9658	5.93 / 9.33
orthogonal	Qwen→BGE	0.5169	0.5168	0.9788	2.39 / 3.45
orthogonal	Qwen→SFR	0.6001	0.5993	0.9828	5.73 / 8.88
orthogonal	SFR→Gemma	0.6032	0.6024	0.9787	1.64 / 2.59
orthogonal	SFR→Qwen	0.6051	0.6046	0.9840	5.77 / 8.58
ridge	BGE→Qwen	0.2343	0.2331	0.9361	7.46 / 9.90
ridge	Gemma→SFR	0.2826	0.2822	0.9419	6.41 / 9.16
ridge	Qwen→BGE	0.4646	0.4642	0.9780	2.46 / 3.65
ridge	Qwen→SFR	0.4714	0.4711	0.9755	5.70 / 9.17
ridge	SFR→Gemma	0.5897	0.5894	0.9846	1.59 / 2.64
ridge	SFR→Qwen	0.5374	0.5371	0.9854	6.13 / 8.86

The retained experiment records also cover ANN200, relevance Recall@100, latency samples, index builds, memory observations, other allocation conditions, native-anchor sensitivity results, and seven-encoder quality references. The latter require all encoder representations per item and are explicitly outside the one-vector-per-document interoperability contract. No supporting STS, clustering, unpaired translation, or compact-student result is claimed as part of this manuscript.

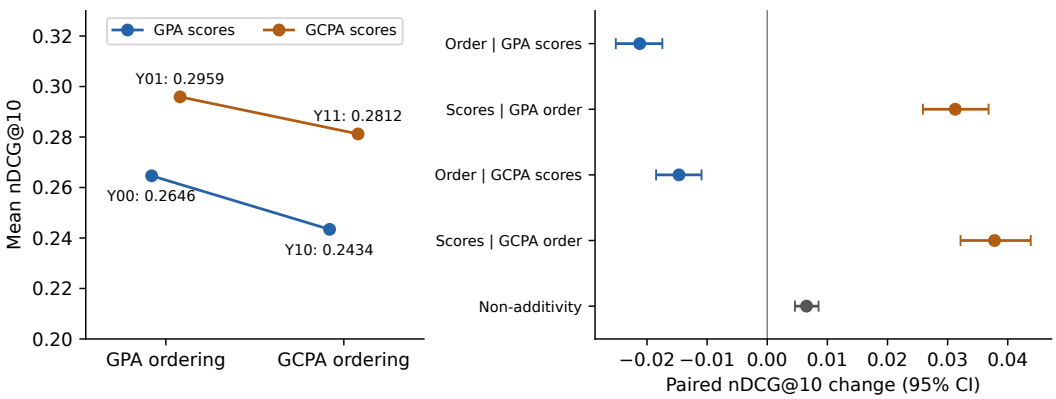


Figure 11: Original row-policy score-order transplant on NFCorpus. Four-arm means and nominal paired nDCG contrast intervals, averaged over seven query encoders and three assignments at fit42. Segments connect measured ordering choices under each donor, not a continuous intervention. The later strict donor-key policy is compared numerically in section J.1.

I SCORE-EXCHANGE DIAGNOSTIC: ORIGINAL PROTOCOL AND DETAILED RESULTS

I.1 PROTOCOL, FOUR ARMS, AND CONDITIONAL CONTRASTS

The follow-up protocol was frozen on September 17, 2026, after the original NFCorpus results and diagnostic hypothesis were exposed. It retains all 323 official test queries, 3,633 documents, seven query encoders, fitting seed42, and assignment seeds42/43/44. Each assignment has exactly 519 documents per source. The four arms yield 84 query-encoder/assignment/arm cells. No new training, neural encoding, ANN construction, dataset acquisition, or calibration selection enters this diagnostic.

The original follow-up audit byte-verified 170 files, confirmed complete 84-cell coverage and passed all 21 identity conditions, with no scientific-grid deviations or numerical failed attempts. We retained checksummed source/protocol identities and intermediate status records to distinguish completed results from earlier partial runs.

This appendix preserves the *original recipient-row tie policy*. The later exact donor-key sensitivity and practical calibration experiments are reported in section J; neither overwrites this original record.

Table 21: All four follow-up arms, with means over seven query encoders and three assignments at fit42. Each donor supplies the complete sorted query-specific score multiset within a shard; it is not the fixed held-anchor affine calibration. Y_{00} and Y_{11} reproduce original mixed GPA/GCPA. Paired contrast intervals below address the comparisons.

Arm	Ordering	Score donor	nDCG@10	Recall@100
Y_{00}	GPA	GPA	0.2646	0.2542
Y_{01}	GPA	GCPA	0.2959	0.3081
Y_{10}	GCPA	GPA	0.2434	0.2387
Y_{11}	GCPA	GCPA	0.2812	0.2952

The predeclared follow-up contrasts are

$$D = Y_{11} - Y_{00}, \qquad O_0 = Y_{10} - Y_{00}, \qquad S_0 = Y_{01} - Y_{00},$$
$$O_1 = Y_{11} - Y_{01}, \qquad S_1 = Y_{11} - Y_{10}, \qquad I = Y_{11} - Y_{10} - Y_{01} + Y_{00}. \tag{18}$$

For each query and each metric, $D = O_0 + S_0 + I$, $O_1 = O_0 + I$, and $S_1 = S_0 + I$. These algebraic identities hold for the constructed intervention; their components are not percentages of a causal training effect. We use the same 10,000 paired query-group draws for every arm, metric contrast, and matched-weight contrast, averaging the 21 conditions within query first. All 323 normalized-query groups are distinct in this panel.

Table 22: All six follow-up contrasts for both metrics, with nominal paired 95% intervals. The positive nDCG non-additivity contrast does not imply non-additivity in Recall: its interval includes zero. No equivalence or multiple-comparison-adjusted discovery claim is made.

Contrast	nDCG@10 [95% CI]	Recall@100 [95% CI]
D	$+0.0166$ $[+0.0095, +0.0238]$	$+0.0410$ $[+0.0339, +0.0479]$
O0	-0.0212 $[-0.0252, -0.0174]$	-0.0155 $[-0.0198, -0.0115]$
S0	$+0.0313$ $[+0.0259, +0.0368]$	$+0.0538$ $[+0.0478, +0.0600]$
O1	-0.0147 $[-0.0185, -0.0109]$	-0.0129 $[-0.0182, -0.0076]$
S1	$+0.0378$ $[+0.0321, +0.0438]$	$+0.0565$ $[+0.0504, +0.0627]$
I	$+0.0065$ $[+0.0046, +0.0085]$	$+0.0026$ $[-0.0005, +0.0058]$

I.2 Matched-weight comparison and heterogeneous outcomes

The homogeneous operand for H_{49} averages the 49 paired GCPA–GPA differences at fit42, including each method's projected diagonals. The H_{42} sensitivity excludes diagonals. Neither uses the original three-fit-seed homogeneous aggregate. With $X = D$ from the mixed arms, the matched contrasts are $J_{49} = X - H_{49}$ and $J_{42} = X - H_{42}$. Identical query IDs and paired draws are retained. Equal encoder/source weights do not make the homogeneous and mixed candidate representations identical, and the resulting contrast is descriptive.

Table 23: Post-hoc matched-weight deployment comparisons, fit 42 only. Both the all-49 primary definition and off-diagonal-42 sensitivity are retained for both metrics. Intervals are nominal paired query-group intervals.

Statistic	nDCG@10 [95% CI]	Recall@100 [95% CI]
X	$+0.0166$ $[+0.0095, +0.0238]$	$+0.0410$ $[+0.0339, +0.0479]$
H49	-0.0301 $[-0.0365, -0.0238]$	-0.0162 $[-0.0211, -0.0112]$
J49	$+0.0467$ $[+0.0401, +0.0536]$	$+0.0572$ $[+0.0506, +0.0639]$
H42	-0.0288 $[-0.0354, -0.0224]$	-0.0149 $[-0.0197, -0.0098]$
J42	$+0.0454$ $[+0.0387, +0.0526]$	$+0.0559$ $[+0.0492, +0.0627]$

Table 24: Descriptive nDCG contrasts by query encoder (averaging three assignments) and by assignment (averaging seven encoders). Total D is negative for Gemma, Jina, and PPLX; Jina's S_0 is slightly negative. O_0 and O_1 have negative means in every row. These rows are not independent inferential samples.

Condition	D	O0	S0	O1	S1	I
Gemma	-0.0144	-0.0241	$+0.0061$	-0.0205	$+0.0097$	$+0.0036$
BGE	$+0.0461$	-0.0204	$+0.0547$	-0.0086	$+0.0665$	$+0.0118$
Arctic	$+0.0111$	-0.0191	$+0.0255$	-0.0145	$+0.0302$	$+0.0047$
Jina	-0.0161	-0.0245	-0.0003	-0.0158	$+0.0084$	$+0.0086$
PPLX	-0.0093	-0.0278	$+0.0107$	-0.0201	$+0.0185$	$+0.0077$
Qwen	$+0.0093$	-0.0214	$+0.0261$	-0.0169	$+0.0307$	$+0.0045$
SFR	$+0.0895$	-0.0112	$+0.0960$	-0.0065	$+0.1008$	$+0.0048$
Assignment 42	$+0.0110$	-0.0234	$+0.0269$	-0.0160	$+0.0344$	$+0.0075$
Assignment 43	$+0.0180$	-0.0218	$+0.0331$	-0.0151	$+0.0398$	$+0.0067$
Assignment 44	$+0.0208$	-0.0184	$+0.0337$	-0.0130	$+0.0392$	$+0.0054$

I.3 SOURCE RECOVERY, EXACT TIES, AND LIMITS OF ISOLATION

Table 25: Source-specific judged-positive recovery, micro within a condition/source then averaged over query encoders and three assignments. This differs from query-macro Recall@100 and from the single-assignment profile in figure 8. All document sources are retained, including SFR's decline. Column labels give ordering/score donor.

Document source	GPA/GPA	GPA/GCPA	GCPA/GPA	GCPA/GCPA
Gemma	0.1250	0.1863	0.1155	0.1782
BGE	0.1471	0.1707	0.1375	0.1627
Arctic	0.0935	0.1369	0.0875	0.1322
Jina	0.1239	0.1891	0.1111	0.1743
PPLX	0.1150	0.1538	0.1048	0.1484
Qwen	0.2256	0.2719	0.2087	0.2534
SFR	0.3601	0.3060	0.3369	0.2861

Across assignments, SFR-source recovery falls from 0.3601 under Y_{00} to 0.3060 under Y_{01} and 0.2861 under Y_{11}. The aggregate score-donation advantage does not give every source a recovery improvement. The saved GPA affine control remains at nDCG 0.2617 versus 0.2646 for Y_{00}; it uses fixed held-anchor statistics rather than query-specific score donation. A successful artificial transplant therefore does not establish that a learned practical calibrator with the same benefit exists.

Table 26: Exact donor-tie audit, summed across query-encoder/assignment/query/source occurrences. Counts are not unique documents. "Changed positions" compares the recipient ordering used in score assignment with the ordering after global-row tie-breaking. Neither identity arm changes positions. No jitter or relaxed identity criterion was used.

Arm	Adjacent ties	Tied occurrences	Changed positions
Y00	33,915	67,830	0
Y01	33,915	67,830	33,616
Y10	33,915	67,830	33,808
Y11	33,915	67,830	0

The original transform preserves the donor's entire within-shard score multiset, but tied donor values can merge distinct recipient ranks. Final recipient-row tie-breaking can then reorder those documents. The 33,616 and 33,808 changed-position counts for Y_{01} and Y_{10} qualify the original language of holding ordering fixed. They do not themselves quantify an effect on nDCG or Recall. The subsequently completed exact policy comparison in section J.1 supplies that measurement without jitter or deletion of tied documents.

The independent follow-up audit checked identity top100 rows, metric reconstructions, 98 homogeneous method/cell panels, 21 saved affine cells, full donor score multisets and factorial identities. Manuscript verification matches original nDCG panels and reconstructs all 30 arm/contrast/matched intervals to absolute tolerance 10^{-12}. We retained the protocol, audit records, all 84 cell summaries and per-query panels. This original follow-up is limited to one exposed domain and one fit. The later cross-domain evidence in section K.2 does not turn the artificial diagnostic into a causal training intervention or cheap calibrator.

J EXACT ORDER PRESERVATION AND QUERY-ADAPTIVE CALIBRATION

J.1 EXACT DONOR-KEY CONSTRUCTION AND POLICY EFFECTS

The separately frozen revision protocol reconstructs all 84 original row-policy cells and evaluates 84 donor-key cells on the same NFCorpus inputs, fit 42 and assignments 42/43/44. For a fixed query/source shard, sort the ordering and donor scores by descending value then ascending global row, yielding $\pi_o(k)$ and $\pi_s(k)$. Attach to the recipient $d = \pi_o(k)$ the unchanged donated score $v_s(k)$ and integer secondary key $\rho(\pi_s(k))$. The final union sorts these tuples as

$$\kappa_{o,s}(q, \pi_o(k)) = \big(- v_s(k), \rho(\pi_s(k)) \big) \tag{19}$$

in ascending lexicographic order and returns the *recipient* document ID. The donor score multiset is byte-identical to that under the original transplant. Equal donor scores have increasing donor-row keys with rank, so the within-shard recipient total order is preserved exactly. For either identity arm, donor and recipient rows coincide, recovering the original global ranking even with cross-source ties. Because documents are disjoint across shards, keys are globally unique; the donor's global representation-source slot sequence is preserved. These are properties of this diagnostic construction, not a theorem about GCPA training.

The independent audit verifies numeric multiset preservation, exact recipient ordering, identity top100 rows, full donor source-slot sequence, unique returned IDs, original/new metric reconstruction, and factorial identities. Exact equality in saved float64 scores defines ties. Adjacent positive gaps at most 10^{-12} are counted separately as near-ties, never rounded or merged for scoring. Both policies and all query/source/assignment breakdowns were retained for verification.

Table 27: Donor-key nDCG means/contrasts and paired differences from the original recipient-row policy, nominal 95% intervals. Six-decimal display can round very small nonzero effects to zero. Identity-arm and total-contrast invariance were checked directly on the unrounded values.

Statistic	Donor-key [95% CI]	Key minus row [95% CI]
Y00	$+0.264634$ $[+0.239445, +0.290199]$	$+0.000000$ $[+0.000000, +0.000000]$
Y01	$+0.295899$ $[+0.267634, +0.324018]$	-0.000000 $[-0.000027, +0.000024]$
Y10	$+0.243412$ $[+0.219170, +0.267532]$	-0.000013 $[-0.000035, +0.000000]$
Y11	$+0.281222$ $[+0.253723, +0.308556]$	$+0.000000$ $[+0.000000, +0.000000]$
D	$+0.016588$ $[+0.009475, +0.023804]$	$+0.000000$ $[+0.000000, +0.000000]$
O0	-0.021222 $[-0.025194, -0.017455]$	-0.000013 $[-0.000035, +0.000000]$
S0	$+0.031265$ $[+0.025888, +0.036805]$	-0.000000 $[-0.000027, +0.000024]$
O1	-0.014677 $[-0.018478, -0.010932]$	$+0.000000$ $[-0.000024, +0.000027]$
S1	$+0.037810$ $[+0.032145, +0.043825]$	$+0.000013$ $[+0.000000, +0.000035]$
I	$+0.006545$ $[+0.004622, +0.008547]$	$+0.000013$ $[-0.000017, +0.000046]$

Table 28: Recall policy sensitivity. Per-query Recall and its contrasts are invariant here despite some changed retrieved grade positions. Recall counts positive hits, not their within-list positions or graded gain. Its non-additivity interval continues to contain zero.

Statistic	Donor-key [95% CI]	Key minus row [95% CI]
Y00	$+0.254247$ $[+0.233270, +0.276131]$	$+0.000000$ $[+0.000000, +0.000000]$
Y01	$+0.308079$ $[+0.282161, +0.334784]$	$+0.000000$ $[+0.000000, +0.000000]$
Y10	$+0.238730$ $[+0.217959, +0.260292]$	$+0.000000$ $[+0.000000, +0.000000]$
Y11	$+0.295210$ $[+0.269487, +0.321543]$	$+0.000000$ $[+0.000000, +0.000000]$
D	$+0.040963$ $[+0.033949, +0.047902]$	$+0.000000$ $[+0.000000, +0.000000]$
O0	-0.015518 $[-0.019777, -0.011450]$	$+0.000000$ $[+0.000000, +0.000000]$
S0	$+0.053832$ $[+0.047828, +0.059981]$	$+0.000000$ $[+0.000000, +0.000000]$
O1	-0.012869 $[-0.018166, -0.007591]$	$+0.000000$ $[+0.000000, +0.000000]$
S1	$+0.056480$ $[+0.050363, +0.062671]$	$+0.000000$ $[+0.000000, +0.000000]$
I	$+0.002649$ $[-0.000477, +0.005752]$	$+0.000000$ $[+0.000000, +0.000000]$

The Y_{01} and Y_{10} policies differ in top10 grades in only five and three query-condition occurrences, respectively (table 29), despite their much larger full-rank position changes. This explains why

rank-displacement counts cannot be read as relevance losses. All 15 audited source/pair combinations in the tie-origin check have identical normalized text and byte-identical raw float32 vectors. These are source/pair combinations, not independent queries or proof about every possible origin of score compression.

Table 29: Original-row versus donor-key policy differences, summed across the 21 conditions and 323 queries. Grade-change columns count affected query-condition occurrences, not changed documents or independent samples. Full changed positions count rank-position occurrences.

Arm	Changed positions	Changed top10 grades	Changed top100 grades
Y00	0	0	0
Y01	33,616	5	35
Y10	33,808	3	22
Y11	0	0	0

Both identity arms and the total difference D remain invariant, as do $X, H_{49}, H_{42}, J_{49}, J_{42}$ when recomputed from the same homogeneous panels. We therefore retain the matched table in table 23. The strict-order control changes the measured non-additivity only slightly in nDCG units; it does not establish equivalence across all possible tie conventions or identify a causal percentage mediated.

J.2 Bank-matched fixed and query-adaptive scoring

We use the r3 fit42 held-anchor membership order, resolving its stored row IDs before accessing arrays. Qbank uses positions $[0, 128)$ and Dbank uses $[128, 640)$, with paired Dbank texts across sources. These are previously used validation anchors, not newly held-out evidence. All bank vectors use the frozen corpus-role GPA feature maps. No retrieval relevance labels select bank rows, rule parameters or checkpoints.

For source b, let $B_b \in \mathbb{R}^{512\times768}$ stack its normalized Dbank rows. Given an actual normalized GPA query row $z_a^q(q)$, C1 computes

$$\begin{aligned} v_{q,b} &= z_a^q(q) B_b^\top, \\ \mu_{q,b} &= \mathrm{mean}(v_{q,b}), \qquad \sigma_{q,b} = \max\{\mathrm{std}_{\mathrm{population}}(v_{q,b}), 10^{-8}\}, \\ s_{C1}(q,d) &= \frac{s_{\mathrm{GPA}}(q,d) - \mu_{q,\alpha(d)}}{\sigma_{q,\alpha(d)}}. \end{aligned} \tag{20}$$

C0 uses mean/population-std over all 128-by-512 similarities from source-a corpus-role Qbank to B_b, fixed per query-encoder/document-source pair. Thus C0 and C1 share Dbank while differing in query conditioning. The old 128-by-128 control is a distinct reference. Computation uses original float32 projections with float64 score/statistics/calibration arithmetic; no floor activations or nonfinite values were recorded.

For each positive affine rule, source-local exact top100 candidates suffice for exact calibrated global top100, with the original global-row tie rule. Full calibrated-scoring checks agree exactly on every NFCorpus query and the first 32 official test queries of each source domain, for every condition and calibrator. All 126 new quality cells and 189 original GPA/GCPA/old-affine reference cells passed independent metric reconstruction. Per-query summaries average the seven encoders and three assignments before 10,000 shared seed 42 group draws. All domains in this original calibration revision are exposed/post-hoc.

Table 30: All predeclared calibration comparisons for both relevance metrics. C1 improves the GPA domain aggregates, but does not erase GCPA gaps across domains or metrics. NFCorpus C1−GCPA nDCG includes zero; its Recall interval is negative. These nominal intervals are not noninferiority/equivalence tests.

Domain	Contrast	nDCG change [95% CI]	Recall change [95% CI]
SciFact	C1-GPA	$+0.0596$ $[+0.0515, +0.0679]$	$+0.0386$ $[+0.0288, +0.0484]$
SciFact	C1-GCPA	-0.0608 $[-0.0726, -0.0485]$	-0.0543 $[-0.0672, -0.0415]$
SciFact	C1-C0	$+0.0771$ $[+0.0622, +0.0923]$	$+0.0467$ $[+0.0338, +0.0600]$
SciFact	C0-GPA	-0.0175 $[-0.0344, -0.0009]$	-0.0080 $[-0.0264, +0.0100]$
FiQA	C1-GPA	$+0.0602$ $[+0.0553, +0.0653]$	$+0.0758$ $[+0.0703, +0.0814]$
FiQA	C1-GCPA	-0.0518 $[-0.0593, -0.0444]$	-0.0804 $[-0.0891, -0.0716]$
FiQA	C1-C0	$+0.0357$ $[+0.0292, +0.0425]$	$+0.0790$ $[+0.0704, +0.0880]$
FiQA	C0-GPA	$+0.0245$ $[+0.0177, +0.0314]$	-0.0032 $[-0.0136, +0.0073]$
NFCorpus	C1-GPA	$+0.0159$ $[+0.0118, +0.0202]$	$+0.0244$ $[+0.0193, +0.0297]$
NFCorpus	C1-GCPA	-0.0007 $[-0.0076, +0.0062]$	-0.0166 $[-0.0225, -0.0106]$
NFCorpus	C1-C0	$+0.0206$ $[+0.0170, +0.0242]$	$+0.0163$ $[+0.0123, +0.0206]$
NFCorpus	C0-GPA	-0.0047 $[-0.0097, +0.0003]$	$+0.0081$ $[+0.0014, +0.0149]$

Table 31: Descriptive C1 differences by query encoder, averaged over assignments. Columns denote C1 minus the named reference, independently for nDCG and Recall. Negative components are retained; these are not independent model samples or multiplicity-adjusted pairwise discoveries.

Domain	Query encoder	ΔGPA nDCG	ΔGCPA nDCG	ΔGPA Recall	ΔGCPA Recall
SciFact	Gemma	+0.0477	−0.0253	−0.0127	+0.0019
SciFact	BGE	+0.0990	−0.0990	+0.1729	−0.1313
SciFact	Arctic	+0.1100	+0.0426	+0.0206	+0.0021
SciFact	Jina	+0.0043	−0.0430	−0.0156	−0.0196
SciFact	PPLX	+0.0180	−0.0264	−0.0029	−0.0150
SciFact	Qwen	+0.0001	−0.1418	−0.0002	−0.1096
SciFact	SFR	+0.1381	−0.1325	+0.1081	−0.1085
FiQA	Gemma	+0.0736	−0.0129	+0.0368	−0.0046
FiQA	BGE	+0.0617	−0.0904	+0.2081	−0.1262
FiQA	Arctic	+0.1125	+0.0298	+0.0648	+0.0347
FiQA	Jina	+0.0223	−0.0566	+0.0046	−0.0551
FiQA	PPLX	+0.0243	−0.0642	−0.0129	−0.0730
FiQA	Qwen	+0.0277	−0.0474	−0.0297	−0.0858
FiQA	SFR	+0.0993	−0.1211	+0.2589	−0.2526
NFCorpus	Gemma	+0.0109	+0.0253	+0.0083	+0.0115
NFCorpus	BGE	+0.0100	−0.0361	+0.0418	−0.0583
NFCorpus	Arctic	+0.0400	+0.0289	+0.0142	+0.0123
NFCorpus	Jina	−0.0036	+0.0125	+0.0069	+0.0145
NFCorpus	PPLX	+0.0124	+0.0217	+0.0137	+0.0061
NFCorpus	Qwen	−0.0054	−0.0147	+0.0169	−0.0074
NFCorpus	SFR	+0.0473	−0.0423	+0.0691	−0.0946

Table 32: All source-specific positive-pair recovery values for GPA/GCPA and the two new controls. Recovery is micro over judged-positive query–document pairs within each condition/source, then averaged over query encoders and assignments. It is not query-macro Recall@100 and is not demographic parity.

Domain	Source	GPA	GCPA	C0	C1
SciFact	Gemma	0.7781	0.9293	0.8533	0.8735
SciFact	BGE	0.8666	0.8865	0.5568	0.7653
SciFact	Arctic	0.6439	0.8787	0.8315	0.6720
SciFact	Jina	0.7696	0.9249	0.9448	0.9234
SciFact	PPLX	0.8180	0.9377	0.9620	0.9106
SciFact	Qwen	0.9140	0.9132	0.9502	0.9524
SciFact	SFR	0.9770	0.9517	0.6494	0.9358
FiQA	Gemma	0.4645	0.6603	0.5259	0.5021
FiQA	BGE	0.4581	0.6034	0.1875	0.4402
FiQA	Arctic	0.3664	0.6059	0.4275	0.3728
FiQA	Jina	0.5445	0.7088	0.8288	0.7558
FiQA	PPLX	0.5253	0.6976	0.7041	0.6525
FiQA	Qwen	0.6856	0.8043	0.8282	0.8128
FiQA	SFR	0.7684	0.8172	0.2826	0.7988
NFCorpus	Gemma	0.1250	0.1782	0.1496	0.1510
NFCorpus	BGE	0.1471	0.1627	0.0363	0.0999
NFCorpus	Arctic	0.0935	0.1322	0.1352	0.0966
NFCorpus	Jina	0.1239	0.1743	0.2395	0.1921
NFCorpus	PPLX	0.1150	0.1484	0.1912	0.1497
NFCorpus	Qwen	0.2256	0.2534	0.3482	0.3588
NFCorpus	SFR	0.3601	0.2861	0.0773	0.2173

J.3 Serving-path cost and ANN fidelity

All systems use four CPU threads, FAISS HNSW M32, efConstruction100 and efSearch400, without parameter search. GPA-single and GCPA-single use their mixed indexes; GPA-shards, C0-shards and C1-shards share seven GPA source-partitioned indexes. Each document contributes one vector. The full-test ANN reference is each system's own exact calibrated top100. Directly measured

Table 33: Original review-revision cached-final-feature benchmark, assignment 42 only. ANN quality/agreement columns average seven query encoders on full test sets. The last columns are medians across seven encoder p50s/p95s, not pooled percentiles. Each distribution has 48 development-panel calls. Projection/correction and neural encoding are excluded. The expanded study is in section K.4.

Domain	System	ANN nDCG	ANN agreement	Median p50 ms	Median p95 ms
SciFact	GPA-single	0.4647	0.9980	0.480	0.541
SciFact	GPA-shards	0.4657	1.0000	1.329	1.347
SciFact	C0-shards	0.4501	1.0000	1.362	1.379
SciFact	C1-shards	0.5276	1.0000	3.211	3.234
SciFact	GCPA-single	0.5858	0.9996	0.464	0.536
FiQA	GPA-single	0.2535	0.9934	0.936	1.213
FiQA	GPA-shards	0.2546	0.9993	5.330	6.504
FiQA	C0-shards	0.2769	0.9997	5.400	6.385
FiQA	C1-shards	0.3163	0.9996	7.717	8.617
FiQA	GCPA-single	0.3655	0.9971	1.019	1.360
NFCorpus	GPA-single	0.2679	0.9930	0.388	0.410
NFCorpus	GPA-shards	0.2687	1.0000	1.067	1.081
NFCorpus	C0-shards	0.2570	1.0000	1.101	1.116
NFCorpus	C1-shards	0.2826	1.0000	2.976	2.992
NFCorpus	GCPA-single	0.2797	0.9976	0.409	0.452

combined time covers calibration, search and merge; marginal percentiles are not added. GCPA batch projection/setup observations do not supply a CPU single-query corrector time.

Dbank payload is 11,010,048 bytes (10.5 MiB); the recorded GPA map payload is 44,826,624 bytes. Original paired-anchor encoding access remains an offline requirement. Serialization, index-build time and process RSS are recorded separately, and every stored vector in 27 indexes was checked against its projection. The separate per-encoder timing table below avoids treating a median of percentiles as a pooled service distribution.

Table 34: Per-encoder combined cached-feature p50/p95 timings. GPA and GCPA columns use single mixed indexes; C1 uses seven shards plus calibration/merge. Source-domain end-to-end neural timings elsewhere in the paper have a different boundary and cannot be substituted here.

Domain	Query encoder	GPA p50/p95 ms	C1 p50/p95 ms	GCPA p50/p95 ms
SciFact	Gemma	0.499/0.606	3.211/3.255	0.453/0.518
SciFact	BGE	0.441/0.512	3.213/3.299	0.505/0.548
SciFact	Arctic	0.504/0.541	3.216/3.245	0.502/0.546
SciFact	Jina	0.480/0.558	3.198/3.217	0.464/0.563
SciFact	PPLX	0.491/0.545	3.209/3.224	0.466/0.536
SciFact	Qwen	0.465/0.505	3.211/3.234	0.453/0.528
SciFact	SFR	0.449/0.479	3.213/3.226	0.448/0.514
FiQA	Gemma	0.953/1.318	7.569/8.864	1.019/1.440
FiQA	BGE	0.936/1.184	8.498/9.206	1.151/1.492
FiQA	Arctic	1.049/1.253	7.832/8.617	1.114/1.423
FiQA	Jina	0.930/1.229	7.538/8.435	0.971/1.291
FiQA	PPLX	0.914/1.213	7.539/8.416	1.010/1.344
FiQA	Qwen	0.972/1.208	7.717/8.336	1.002/1.214
FiQA	SFR	0.862/1.064	8.409/9.064	1.034/1.360
NFCorpus	Gemma	0.397/0.425	2.979/3.013	0.409/0.439
NFCorpus	BGE	0.373/0.394	2.972/2.983	0.416/0.465
NFCorpus	Arctic	0.406/0.431	2.976/2.992	0.411/0.455
NFCorpus	Jina	0.388/0.410	2.981/3.020	0.399/0.449
NFCorpus	PPLX	0.390/0.415	2.980/2.997	0.409/0.452
NFCorpus	Qwen	0.388/0.405	2.971/2.984	0.402/0.452
NFCorpus	SFR	0.374/0.387	2.971/2.989	0.420/0.457

The original tie/calibration revision audit records 836 remotely hashed files and 803 locally verified mirrors. Verification used retained protocols, audits, summaries, per-query metrics and timing records; manuscript checks reproduce 124 intervals and 420 timing distributions. Training-factor and

historical-NQ phases were not executed in that revision. The subsequent completed historical-NQ extension and other new evidence are reported separately below.

K GENERALIZATION, CALIBRATION SENSITIVITY AND EXPANDED MEASUREMENTS

The following experiments ask whether the reversal and its score/order pattern extend to a new domain, whether a simple calibration modification helps, and whether the serving conclusions survive broader measurement. We report the prospective Android test first, followed by complete cross-domain diagnostics, calibration sensitivity, expanded serving measurements, historical NQ, and document throughput. These experiments add 1,022 exact-quality conditions: 616 Android, 168 SciFact/FiQA transplant, 126 variance-sensitivity and 112 historical-NQ conditions. Conditions are averaged within query; they are not independent statistical samples. We retained each condition and its per-query, source and encoder summaries separately from the earlier experiments.

K.1 PROSPECTIVE ANDROID PROTOCOL AND ADJUSTED DECISIONS

Android tests predictions made after the original-domain analyses on a previously unexamined retrieval condition. We preselected the Android subcorpus from the official BEIR CQADupStack archive (Hoogeveen et al., 2015; Thakur et al., 2021), freezing archive checksums, file identities, title/body assembly, row IDs and relevance judgments. All 22,998 documents and 699 test queries are retained, with no exact query-ID self matches in the corpus. The exposure audit found no prior project outcome for this condition and no normalized exact text overlap with 28,928 used anchor-pool rows. These checks do not establish absence from encoder pretraining or disjoint original document ancestry. The source is Stack Exchange content under CC BY-SA; the intake does not recover each post's original license-version history.

The three frozen GPA/GCPA fits at seeds 42/43/44 are reused, with three shared balanced assignments per fit and all seven query encoders. Each document supplies only its assigned source vector. Native same-model references are not duplicated across fits. The 616 cells comprise $2 \times 3 \times 49 = 294$ homogeneous conditions, $3 \times 3 \times 7 \times 4 = 252$ donor-key mixed conditions, 63 C1 conditions and seven native references. Native/projected diagonals remain separate from the off-diagonal mean.

The encoder recipe has one important provenance qualification. Primary model revisions and role settings were retained, and newly resolved tokenizer/remote-code dependencies were pinned before outcomes, but equality to every historical dependency could not be certified, particularly for Jina. We accepted this qualified condition before seeing Android quality results. A tokenizer-serialization mismatch in two initial jobs was resolved by comparing matched states: all 32 pilot embeddings were byte-identical before full encoding resumed. Qwen/SFR use the fixed instruction "Given an Android search query, retrieve questions that express the same information need." Other models retain their recorded query/document roles. The experiment thus tests the frozen, realized recipe rather than claiming exact reproduction of every historical dependency.

Query groups are connected components of the typed query/eligible-positive-document bipartite graph, additionally joining identical normalized query texts. Component identities hash sorted query IDs; all 699 realized groups contain one query. Conditions are averaged within query before 50,000 shared seed42 group-bootstrap draws. Nominal intervals use the 2.5th/97.5th percentiles; the six primary directions additionally use one-sided Bonferroni bounds at $\alpha/6$, with linear quantile interpolation. These are approximate bootstrap bounds conditional on the fixed cohort, corpus and assignments, not finite-sample coverage guarantees. All six pass; secondary quantities are not promoted into this family.

Android C1 uses the original 512-anchor rule, with bank positions $[128, 640)$ in each fit's held membership. Banks 43/44 are derived from existing encoded anchors; bank42 matches the original bank. The three-fit estimand differs from the original fit42-only mixed results. A separately retained fit42 Android panel gives $X = +0.1076$ $[+0.0984, +0.1168]$; it is secondary and not selected in place of the three-fit result. Verification retained complete homogeneous/projected/native references and fit/encoder/assignment breakdowns.

Table 35: Android's six predeclared nDCG directions. The adjusted bound is upper for a negative claim and lower for a positive claim; each is strictly on its declared side of zero. Nominal intervals are displayed separately. $D = X$.

Direction	nDCG change [nominal 95% CI]	Adjusted bound
$H42 < 0$	-0.0162 $[-0.0212, -0.0112]$	-0.010092
$D > 0$	$+0.1027$ $[+0.0944, +0.1111]$	$+0.092566$
$S0 > 0$	$+0.1058$ $[+0.0980, +0.1136]$	$+0.096398$
$S1 > 0$	$+0.1104$ $[+0.1025, +0.1183]$	$+0.100705$
$O0 < 0$	-0.0076 $[-0.0093, -0.0060]$	-0.005650
$O1 < 0$	-0.0031 $[-0.0053, -0.0008]$	-0.000391

K.2 Diagnostic transfer and complete metric contrasts

We repeat the order-preserving score exchange on SciFact and FiQA to determine whether the original NFCorpus diagnostic is domain-specific. Both use the unchanged full test sets, fit seed 42, seven query encoders, three original assignments and all four combinations of order and scores, for 168 conditions. Donations use every eligible score rather than a truncated candidate list. Independent checks verify the original GPA/GCPA top-100 rankings, exact recipient order, donated score multisets, source-slot sequences, unique returned IDs and the contrast identities. FiQA's 38 empty documents and duplicate-text IDs remain eligible. A deadline-limited worker completed 73 of 84 FiQA conditions; a versioned sorting optimization reproduced all 73 predictions exactly before finishing the unchanged grid. This repair did not alter the scientific comparison.

The complete results below show where the shared pattern is uncertain. NFCorpus comes from the earlier donor-key experiment; SciFact and FiQA are post-hoc extensions; Android is the prospective condition with the recipe qualification above. All intervals here are nominal, including Android's secondary quantities. Original-domain intervals use 10,000 normalized-query-group draws. FiQA's O_1 and SciFact's interaction I include zero in nDCG, so negative ordering point estimates and non-additivity should not be treated as universal findings. Per-query, encoder, assignment and source-recovery breakdowns are retained alongside the domain means.

Table 36: Complete score-exchange results: four arms and six contrasts for both metrics. All intervals are nominal paired 95% intervals; the adjusted Android primary tests are reported separately.

Domain	Contrast	nDCG [95% CI]	Recall [95% CI]
SciFact	Y00	$+0.4613$ $[+0.4336, +0.4884]$	$+0.8206$ $[+0.7977, +0.8423]$
SciFact	Y01	$+0.5923$ $[+0.5577, +0.6255]$	$+0.9278$ $[+0.9082, +0.9457]$
SciFact	Y10	$+0.4484$ $[+0.4198, +0.4766]$	$+0.8003$ $[+0.7749, +0.8241]$
SciFact	Y11	$+0.5817$ $[+0.5455, +0.6162]$	$+0.9135$ $[+0.8917, +0.9331]$
SciFact	D	$+0.1204$ $[+0.1058, +0.1349]$	$+0.0929$ $[+0.0802, +0.1060]$
SciFact	S0	$+0.1310$ $[+0.1185, +0.1436]$	$+0.1071$ $[+0.0958, +0.1187]$
SciFact	S1	$+0.1332$ $[+0.1205, +0.1460]$	$+0.1132$ $[+0.1009, +0.1263]$
SciFact	O0	-0.0129 $[-0.0166, -0.0093]$	-0.0203 $[-0.0271, -0.0142]$
SciFact	O1	-0.0106 $[-0.0150, -0.0065]$	-0.0143 $[-0.0225, -0.0065]$
SciFact	I	$+0.0022$ $[-0.0001, +0.0044]$	$+0.0061$ $[-0.0002, +0.0123]$
FiQA	Y00	$+0.2478$ $[+0.2337, +0.2624]$	$+0.5714$ $[+0.5540, +0.5885]$
FiQA	Y01	$+0.3621$ $[+0.3420, +0.3828]$	$+0.7342$ $[+0.7143, +0.7535]$
FiQA	Y10	$+0.2382$ $[+0.2243, +0.2528]$	$+0.5508$ $[+0.5328, +0.5687]$
FiQA	Y11	$+0.3598$ $[+0.3395, +0.3808]$	$+0.7276$ $[+0.7066, +0.7483]$
FiQA	D	$+0.1120$ $[+0.1031, +0.1210]$	$+0.1562$ $[+0.1471, +0.1653]$
FiQA	S0	$+0.1143$ $[+0.1061, +0.1225]$	$+0.1628$ $[+0.1559, +0.1697]$
FiQA	S1	$+0.1216$ $[+0.1130, +0.1300]$	$+0.1768$ $[+0.1689, +0.1848]$
FiQA	O0	-0.0095 $[-0.0115, -0.0076]$	-0.0206 $[-0.0248, -0.0165]$
FiQA	O1	-0.0023 $[-0.0049, +0.0002]$	-0.0066 $[-0.0126, -0.0007]$
FiQA	I	$+0.0072$ $[+0.0054, +0.0090]$	$+0.0140$ $[+0.0093, +0.0185]$
NFCorpus	Y00	$+0.2646$ $[+0.2394, +0.2902]$	$+0.2542$ $[+0.2333, +0.2761]$
NFCorpus	Y01	$+0.2959$ $[+0.2676, +0.3240]$	$+0.3081$ $[+0.2822, +0.3348]$
NFCorpus	Y10	$+0.2434$ $[+0.2192, +0.2675]$	$+0.2387$ $[+0.2180, +0.2603]$
NFCorpus	Y11	$+0.2812$ $[+0.2537, +0.3086]$	$+0.2952$ $[+0.2695, +0.3215]$
NFCorpus	D	$+0.0166$ $[+0.0095, +0.0238]$	$+0.0410$ $[+0.0339, +0.0479]$

Table 36 (continued)

Domain	Contrast	nDCG [95% CI]	Recall [95% CI]
NFCorpus	S0	$+0.0313$ [$+0.0259$, $+0.0368$]	$+0.0538$ [$+0.0478$, $+0.0600$]
NFCorpus	S1	$+0.0378$ [$+0.0321$, $+0.0438$]	$+0.0565$ [$+0.0504$, $+0.0627$]
NFCorpus	O0	-0.0212 [-0.0252, -0.0175]	-0.0155 [-0.0198, -0.0115]
NFCorpus	O1	-0.0147 [-0.0185, -0.0109]	-0.0129 [-0.0182, -0.0076]
NFCorpus	I	$+0.0065$ [$+0.0046$, $+0.0085$]	$+0.0026$ [-0.0005, $+0.0058$]
Android	Y00	$+0.2505$ [$+0.2370$, $+0.2641$]	$+0.5894$ [$+0.5710$, $+0.6074$]
Android	Y01	$+0.3563$ [$+0.3367$, $+0.3761$]	$+0.7656$ [$+0.7452$, $+0.7856$]
Android	Y10	$+0.2429$ [$+0.2294$, $+0.2564$]	$+0.5728$ [$+0.5541$, $+0.5911$]
Android	Y11	$+0.3532$ [$+0.3334$, $+0.3731$]	$+0.7597$ [$+0.7384$, $+0.7803$]
Android	D	$+0.1027$ [$+0.0944$, $+0.1111$]	$+0.1703$ [$+0.1612$, $+0.1793$]
Android	S0	$+0.1058$ [$+0.0980$, $+0.1136$]	$+0.1763$ [$+0.1687$, $+0.1838$]
Android	S1	$+0.1104$ [$+0.1025$, $+0.1183$]	$+0.1869$ [$+0.1785$, $+0.1951$]
Android	O0	-0.0076 [-0.0093, -0.0060]	-0.0166 [-0.0198, -0.0134]
Android	O1	-0.0031 [-0.0053, -0.0008]	-0.0060 [-0.0107, -0.0014]
Android	I	$+0.0046$ [$+0.0032$, $+0.0059$]	$+0.0106$ [$+0.0072$, $+0.0139$]

K.3 Variance shrinkage and bank diagnostics

The calibration sensitivity asks whether modest smoothing of reference-bank variances improves C1. For query/source population bank variance v_{qb} and mean μ_{qb}, the prespecified rule C1-V uses

$$\sigma^V_{qb} = \max\left\{\sqrt{0.9v_{qb} + 0.1\,\mathrm{median}_j\, v_{qj}}, 10^{-8}\right\}, \qquad s_{C1\text{-}V}(q,d) = \frac{s_{\mathrm{GPA}}(q,d) - \mu_{q,\alpha(d)}}{\sigma^V_{q,\alpha(d)}}. \tag{21}$$

Means, bank membership, eligibility, GPA coordinates and source-local merging remain fixed. There is no weight/floor/temperature search, relevance-based selection or new fitting. Its positive affine source transforms are checked against full calibrated scoring, and zero-shrinkage replay reproduces C1. The 126 cells comprise 63 across the three old domains at fit42 and 63 across three Android fits. These secondary intervals use the corresponding domain's paired draws; Android's primary six-direction family is unchanged.

Table 37: Fixed C1-V sensitivity. All nDCG point differences from C1 are negative; SciFact's interval includes zero. Both metrics are retained. This tests one rule, not the sufficiency of all practical calibration.

Domain	C1-V nDCG	ΔC1 nDCG [95% CI]	ΔC1 Recall [95% CI]
SciFact	0.5197	-0.0012 [-0.0030, $+0.0007$]	$+0.0036$ [$+0.0015$, $+0.0060$]
FiQA	0.3057	-0.0022 [-0.0034, -0.0011]	$+0.0048$ [$+0.0031$, $+0.0064$]
NFCorpus	0.2795	-0.0011 [-0.0016, -0.0006]	-0.0004 [-0.0012, $+0.0004$]
Android	0.2946	-0.0029 [-0.0037, -0.0021]	$+0.0029$ [$+0.0017$, $+0.0042$]

We also examine whether the estimated calibration coefficients vary across reference texts. Splitting the same paired 512-anchor bank into even and odd rows gives two 256-row estimates; these are diagnostics, not additional retrieval systems. We record absolute log standard-deviation ratios and standardized mean differences, alongside C1's gains or losses against GPA and GCPA in all 28 domain/query-encoder panels. Outcome-independent quintile bins summarize the associations. The complete records retain un-floored coefficients, inverse scales, offsets and source quantiles. These checks help distinguish a quality gap from numerical instability: for example, SciFact SFR's C1 result is 0.1325 below GCPA but 0.1381 above GPA. An association with bank coefficients, or a failed smoothing rule, does not identify a causal source of that gap.

K.4 Expanded serving measurement

The expanded serving study addresses two limitations of the earlier timing panel: its small query set and its exclusion of query projection/correction. We reuse the fit-seed-42, assignment-seed-42 HNSW indexes ($M = 32$, efConstruction100, efSearch400), four total CPU threads and serial single-query requests. Five fresh benchmark processes each use every official test query, giving 1,500/3,240/1,615

Table 38: Bank-coefficient diagnostics. Minimum std is across all queries/query encoders/document sources/fits. The range spans source-condition medians of $|\log(\sigma_{\text{even}}/\sigma_{\text{odd}})|$, not pooled percentiles or a confidence interval. There are 49 source conditions per original domain and 147 on Android. No full-bank floor event occurs; this does not exclude bank/domain mismatch or estimation sensitivity.

Domain	Minimum σ	Half-bank median range	Floor events
SciFact	0.0293	0.0382–0.0979	0
FiQA	0.0272	0.0367–0.0667	0
NFCorpus	0.0265	0.0381–0.0858	0
Android	0.0273	0.0176–0.0902	0

timed calls per encoder/system/boundary on SciFact/FiQA/NFCorpus, but only 300/648/323 distinct queries. Each panel begins with 32 untimed warmups. Query order is paired across systems, pass seeds are 42–46, and system/boundary orders are fixed before timing. No slow calls are removed. Isolated timing nodes and recorded thread pools, runtimes and pass variation support repeatability.

The final-feature boundary excludes preparation uniformly. The native-output boundary begins from the same encoder's cached raw query vector and includes normalization, projection, the GCPA CPU corrector or C1 statistics, search and merging. Corrector timing is a subset of preparation, not an additional disjoint component. Both total paths are directly clocked; marginal percentiles are not added. Neither boundary includes neural text encoding, network or queuing. All seven sequential shard searches are already inside the C1 timer; per-query-encoder panels are not per-shard timers.

Table 39: All 24 pooled serving distributions. "Native" means cached native encoder output, not neural text encoding. Raw calls are pooled with uniform query-encoder/query traffic. Verification additionally used per-encoder and per-pass distributions, correction-only times, ANN quality and exact-agreement references.

Domain	System	Boundary	Pooled p50/p95 ms
SciFact	GPA-single	final	0.471/0.523
SciFact	GPA-single	native	0.704/0.821
SciFact	GCPA-single	final	0.482/0.539
SciFact	GCPA-single	native	1.100/1.248
SciFact	GPA-shards	final	1.322/1.359
SciFact	GPA-shards	native	1.659/1.807
SciFact	C1-shards	final	3.219/3.269
SciFact	C1-shards	native	3.674/3.881
FiQA	GPA-single	final	0.956/1.192
FiQA	GPA-single	native	1.342/1.630
FiQA	GCPA-single	final	1.010/1.282
FiQA	GCPA-single	native	1.843/2.180
FiQA	GPA-shards	final	5.635/6.681
FiQA	GPA-shards	native	6.618/7.820
FiQA	C1-shards	final	7.927/9.033
FiQA	C1-shards	native	8.837/10.085
NFCorpus	GPA-single	final	0.387/0.419
NFCorpus	GPA-single	native	0.599/0.738
NFCorpus	GCPA-single	final	0.414/0.460
NFCorpus	GCPA-single	native	1.015/1.172
NFCorpus	GPA-shards	final	1.064/1.086
NFCorpus	GPA-shards	native	1.370/1.528
NFCorpus	C1-shards	final	2.960/3.016
NFCorpus	C1-shards	native	3.368/3.605

All 168 panels have five passes, totaling 355,880 calls. Numerical feature-path parity and each served path's own exact top100 are checked before comparing ANN quality. The assignment42 quality differs from three-assignment exact means. Repeated calls are not independent new queries; these empirical percentiles and pass variation characterize warm local repeatability, not production tail-latency coverage. The older 48-call neural-service example remains a distinct historical observation.

K.5 FULL HISTORICAL-NQ MATCHED REVERSAL

The NQ extension asks whether both sides of the reversal occur in a much larger index. It retains all 2,681,468 documents and 3,452 queries, the frozen historical GPA maps and cleaned historical paired anchors. Because a corrector depends on its input coordinates, we train and validation-select one seed-42 GCPA corrector in that historical GPA basis using the original declared recipe. Reusing the fresh-domain corrector in a different basis would not be a matched comparison. No neural encoder or GPA map is refitted. Historical dependency provenance remains incomplete, so this experiment concerns the stored vectors rather than their reproduction by current encoder releases.

The grid contains 98 homogeneous cells (2×49 pairs) and 14 mixed cells (two methods, seven query encoders), with one balanced assignment42. It therefore tests both sides of the reversal, unlike a mixed-only scale comparison. Pair/encoder averages are formed within query before 10,000 shared seed42 normalized-query-group draws. The all49 homogeneous mean includes projected diagonals; off-diagonal42 remains separately reported. All NQ intervals are nominal and post-hoc.

Table 40: Full historical-NQ matched deployment contrasts. Negative H_{42} and positive X both exclude zero; positive J alone would not establish the sign reversal.

Contrast	nDCG change [95% CI]	Recall change [95% CI]
H42	-0.0501 $[-0.0533, -0.0467]$	-0.0310 $[-0.0343, -0.0278]$
H49	-0.0499 $[-0.0530, -0.0467]$	-0.0301 $[-0.0330, -0.0270]$
X	$+0.1031$ $[+0.0984, +0.1079]$	$+0.1252$ $[+0.1191, +0.1311]$
J42	$+0.1532$ $[+0.1484, +0.1579]$	$+0.1563$ $[+0.1509, +0.1616]$
J49	$+0.1530$ $[+0.1483, +0.1577]$	$+0.1553$ $[+0.1499, +0.1606]$

One new mixed GCPA HNSW index is checked against every stored projected vector; the compatible original GPA index is reused. All 14 mixed conditions are evaluated at efSearch200/400, yielding 28 ANN-quality conditions relative to each method's own exact ranking. At efSearch400, timing uses all queries over three fresh processes, 32 warmups and 144,984 calls in total, with raw query/order/timestamp/returned-ID provenance. An initial timing version lacked required metadata; it is preserved alongside the protocol-specified complete repetition, rather than selecting a favorable duration. The original six-pair native/direct scale table remains historical, not the operand of the new all-pair mean.

Table 41: Full-NQ mixed exact/ANN nDCG and top100 agreement with each method's own exact ranking, averaged over all seven query encoders. Alignment quality and approximation fidelity remain separate quantities.

Method	efSearch	Exact nDCG	ANN nDCG	Own-exact agreement
GPA	200	0.2601	0.2555	0.9362
GPA	400	0.2601	0.2576	0.9658
GCPA	200	0.3633	0.3592	0.9674
GCPA	400	0.3633	0.3608	0.9865

Table 42: Full-NQ mixed search-only timing, cached final features, efSearch400. Query projection/correction and neural encoding are excluded. These are uniform-traffic raw-call pooled percentiles, not the original 20-call panel.

Method	Cached-feature p50/p95 ms	Calls
GPA	2.005/2.683	72,492
GCPA	1.812/2.498	72,492

NQ supplies a large-index instance of the reversal, not a corpus-size causal effect. Domain, vector recipe and fit provenance differ from fresh-domain conditions. No NQ transplant or C1 experiment is claimed.

K.6 Controlled document-throughput sensitivity

Document encoding is a major component of migration cost, but its throughput can depend on batch size. We therefore encode a fixed hash-selected panel of 1,024 FiQA documents with Gemma and SFR at batch sizes 1, 8 and 32, using three passes and the original text, roles, precision and token caps. The resulting 18,432 document presentations are repeated measurements, not unique documents; existing caches remain unchanged. The harness records tokenization, transfer, synchronized inference, normalization/output copying, remaining processing, total time, padding and memory, as well as panel versus full-corpus token-length coverage.

Table 43: Measured fixed-panel throughput, all six encoder/batch conditions. Larger batches improve Gemma here; SFR batch32 is slower than batch8. Synchronization and instrumentation are part of this harness.

Encoder	Batch	Documents/s	Tokens/s
Gemma	1	35.79	6372.27
Gemma	8	246.18	43830.54
Gemma	32	339.12	60378.21
SFR	1	58.87	10714.49
SFR	8	68.20	12412.70
SFR	32	44.05	8017.56

Batch throughput is therefore not assumed monotone or transferable across encoders. These measurements do not establish full-FiQA/NQ migration runtimes or justify multiplying a selected best panel rate into a production forecast. The original illustrative microbatch extrapolations remain identified as estimates.

K.7 Closest-work access and evaluation contracts

Vejendla (2025) fit paired old/new document vectors to query-to-legacy orthogonal, low-rank affine and residual-MLP adapters. Their reported text-retrieval ground truth is the new model's exhaustive neighbors, rather than human relevance judgments; they explicitly discuss migration, mixed states and operational costs. Several concrete formulas assume equal widths, so unequal-width reproduction requires implementation clarification. Our existing direct orthogonal/ridge comparators are not claimed to reproduce every Drift-Adapter arm.

Maekawa et al. (2026) introduce *Efficient Retrieval Adapter* (ERA): unlabeled target-corpus alignment followed by relevance-supervised adaptation of a rectangular linear query map, preserving backbones and the document index. Their alignment-only variant does not require the supervised stage, but still has a different target-corpus/data-selection contract from frozen Dolma maps and the held-anchor C1 bank. Sun et al. (2024)'s MAIR supplies diverse instructed-retrieval tasks and document collections; it is not itself a mixed-encoder assignment protocol. No ERA or Drift-Adapter numerical result is claimed here, and published scores under different access/label/metric budgets are not inserted into our comparisons.

K.8 Evidence, repairs and reconstruction

The completed audit binds 3,296 byte-verified local artifact mirrors. Independent experimental audits reconstruct metrics from official judgments, rebuild Android query groups, verify exact condition sets and preserve source/input hashes. Technical attempts include the tokenizer-state repair, FiQA deadline continuation, two NQ metadata-write repairs with unchanged predictions, and the complete NQ timing repetition. All attempts count toward the resource ledger; none is a scientific null or an outcome-based change of methods. Parallel ordinary workers and isolated timing workers are documented execution amendments, distinct from scientific protocol choices. During manuscript preparation, a separate reconstruction reproduced 166 additional nominal intervals, Android's six adjusted decisions and 355,880 combined timing calls from retained query-level records. Source/receipt hashes and compact numerical records preserve traceability separately from large vectors, model weights and indexes.