BETTER CONTRAST DECISIONS, WORSE RANKING? HOW VALIDATION SELECTS SCORING FUNCTIONS FOR TEXT EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

A scoring function can rank a passage above its negated counterpart yet place it below other documents in a collection. We study how to select functions that improve this contrast while retaining the benchmark answer's ranking performance. Using frozen text embeddings, we hold each candidate bank fixed within a selection experiment and vary the documents and criteria used for validation. On NevIR, inspection after selection finds candidates that improve contrast accuracy with no decrease in measured test ranking. Yet contrast-only validation repeatedly selects functions with substantial ranking losses. Including retrieved competitors and prioritizing ranking quality reduces these losses while retaining contrast gains. Competitor identity matters even at a fixed list length, and more labeled contrasts do not replace missing background comparisons. The same selected set can also improve average fixed-list reranking while lowering average full-corpus ranking. On NFCorpus, the trained candidates offer only small gains in distinguishing relevance grades, leaving little improvement for a selector to recover. These results distinguish gains available among the trained candidates from gains recovered by validation, and show why a selected scoring function must be evaluated in its intended retrieval setting.

1 INTRODUCTION

Embedding-based retrievers must distinguish passages that share words but express opposite meanings. Consider two passages: one says a service offers overseas delivery; the other says it does not. The first should rank higher for a query seeking overseas delivery, and the second for a query excluding it. NevIR tests these paired decisions, which many dense retrievers fail to make correctly (Weller et al., 2024; van den Elsen et al., 2025).

A scoring function applied to stored embeddings, called a *readout*, can improve these decisions without changing the encoder (Cao, 2025). But the answer may still rank below other documents. Our goal is therefore joint: improve contrast accuracy while limiting loss in corpus ranking. Maximizing contrast accuracy alone need not meet this goal.

Research question. Which validation comparisons identify candidates for this joint goal? A weak selected result may reflect limited gains among the candidates, or a missed useful alternative. Holding the candidate bank fixed within a selection comparison separates these explanations.

Training creates a candidate bank; validation chooses one member; test evaluation measures its outcome. Within a selection comparison, we freeze that bank and change only the validation documents and selection criterion, so the candidate weights and their query–document scores do not change across the compared selectors and views. Each rule selects once for all test queries. We inspect the other candidates' test results afterward to diagnose missed opportunities, without using that inspection to make the validation choice.

Contribution. The fixed-bank design separates what training makes available from what validation recovers. Figure 1 shows why this distinction matters: post-selection test inspection finds a candidate with substantial contrast gain and nonnegative measured ranking change in every NevIR bank, yet

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.

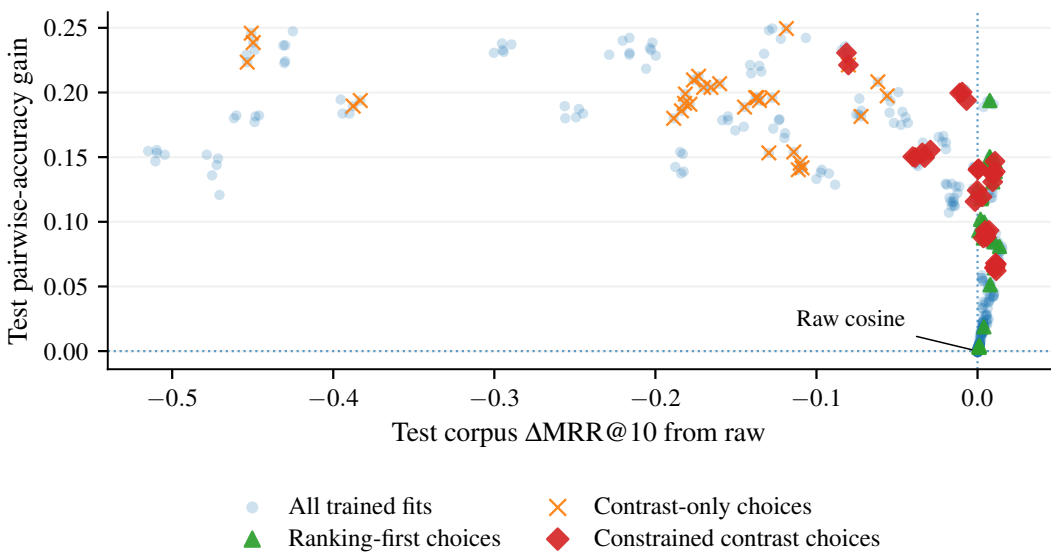


Figure 1: **Useful alternatives can be present in a candidate bank yet not be selected by validation.** Small circles show 480 trained readouts: 16 per bank across six encoders and five seeds. Larger crosses, triangles, and diamonds mark contrast-only, ranking-first, and constrained-contrast choices (30 per rule). Within each bank, rules choose among the same 16 fits and raw cosine using identical raw top-10 validation lists and all 225 pairs. Horizontal and vertical positions give test full-corpus MRR@10 change and pairwise-accuracy gain relative to raw at $(0, 0)$. Above-left points improve contrast but lose ranking; above-right points improve both. Contrast-only favors larger contrast gains with substantial ranking losses; ranking-first retains gains near the no-loss line. Test coordinates are inspected after validation selection, never used to make it. Overlap can hide fits; Tables 1 and 6 give selected outcomes and within-bank availability.

contrast-only validation selects candidates with large ranking losses. Those losses are therefore not forced by the absence of a useful trained alternative in the observed candidate bank.

Competitor identity also affects selection at a fixed list length, with effects that depend on competitor count. The same selected set can improve average fixed-list reranking while lowering average full-corpus ranking. NFCorpus complements NevIR: its trained candidates offer only small grade-preference gains, leaving little additional gain for selection to recover.

2 Related Work

Ranking evidence and model selection. Sampled metrics can reverse system comparisons even in expectation (Krichene & Rendle, 2020). Similar validation scores can hide differences in later use (D'Amour et al., 2022), and compositional training can reduce retrieval generalization (Ralev et al., 2026). We isolate selection from training: which validation views and objectives recover useful choices from the same fitted candidates?

Information accessible to a readout. Probing tests which information can be recovered from representations (Hewitt & Liang, 2019; Ravichander et al., 2021; Belinkov, 2022). Interventions instead modify components and measure behavioral changes (Ravfogel et al., 2020; Elazar et al., 2021; Belrose et al., 2023). LABCLIP aligns frozen embeddings to improve attribute binding (Koishigarina et al., 2025); retrieval studies likewise analyze existing vectors (Goldfarb-Tarrant et al., 2024; Reichman & Heck, 2024; Ram et al., 2023). Our recovery experiments test contrast gains on stored vectors (Appendix G); selection asks whether validation also retains corpus ranking.

Changing retrieval scores. Coordinate reweighting, query-vector updates, and relation-specific scores modify decisions made from embeddings (Cao, 2025; Lee et al., 2026; Yin et al., 2026; Ali et al., 2026). Low-rank ranking metrics are also established tools (Lim & Lanckriet, 2014). We use

one low-rank family with different preservation weights to form candidates with different contrast gains and ranking costs.

3 Problem and experimental design

3.1 Fixed vectors, contrast decisions, and answer ranks

We encode each text once and keep its normalized vector fixed. For encoder e with output dimension D, the reference scoring function uses cosine similarity:

$$z(x) = \frac{e(x)}{\|e(x)\|_2}, \qquad s_{h_0}(q,d) \equiv s_0(q,d) = z(q)^\top z(d). \tag{1}$$

A readout h changes the score $s_h(q,d)$ computed from these stored vectors. We study BGE-M3, BGE-large, multilingual E5, mGTE, and Qwen3-Embedding at 0.6B and 4B (Chen et al., 2024; Wang et al., 2022; Zhang et al., 2024; 2025b).

NevIR has 948 training, 225 validation, and 1,383 test document pairs. Each pair has two opposing queries: d_i answers q_i, and the other document is its *twin*. The answer–twin score margins are:

$$m_1(h) = s_h(q_1,d_1) - s_h(q_1,d_2), \qquad m_2(h) = s_h(q_2,d_2) - s_h(q_2,d_1). \tag{2}$$

Pairwise accuracy $A(h)$ is the fraction of pairs for which both margins are positive. Both query decisions must be correct; independent random choices therefore give accuracy 0.25.

Corpus evaluation follows the same designated answer among pooled documents: 1,761 unique training and validation documents for validation, and 2,081 for 2,766 test queries. Let d_q^* be the designated answer, d_q^{tw} its twin, Q the query set, and $\mathcal{D} = (D_q)_{q\in Q}$ the documents scored for each query. Rank counts documents scoring strictly above the answer:

$$r_h(q;D_q) = \begin{cases} 1 + \displaystyle\sum_{d\in D_q\setminus\{d_q^*\}} \mathbf{1}[s_h(q,d) > s_h(q,d_q^*)], & d_q^* \in D_q, \\ \infty, & d_q^* \notin D_q, \end{cases}$$
$$R_{Q,\mathcal{D}}(h) = \frac{1}{|Q|}\sum_{q\in Q} f_{10}(r_h(q;D_q)), \qquad f_{10}(r) = \frac{\mathbf{1}[r\le 10]}{r}. \tag{3}$$

We set $f_{10}(\infty) = 0$. These metrics answer different questions. Pairwise accuracy asks whether each answer beats its twin; corpus MRR asks how many documents outrank the designated answer. Ranks 1, 2, and 5 contribute 1, $1/2$, and $1/5$; an absent answer or one below rank 10 contributes zero. NevIR labels the paired answer, not every pooled document's relevance. Corpus changes therefore track designated-answer ranks.

3.2 Training a fixed bank of readouts

We learn a low-rank residual change to each normalized vector:

$$T_{U,V}(z) = \text{norm}(z + UV^\top z), \qquad U,V \in \mathbb{R}^{D\times r}. \tag{4}$$

Here $\text{norm}(v) = v/\|v\|_2$. For $h = (U,V)$, we transform both query and document vectors and take their inner product: $s_h(q,d) = T_{U,V}(z(q))^\top T_{U,V}(z(d))$. The matrix $V^\top$ reads combinations of input coordinates, and U maps them back to the original space. The output dimension remains D; rank r limits the update.

The pairwise loss averages $[\mu - m_1(h)]_+ + [\mu - m_2(h)]_+$, with $\mu = 0.05$ and $[u]_+ = \max(u,0)$. We sample 16 training documents per pair, excluding the answer and twin. We center each query's background scores, then penalize squared changes from raw, divided by the pooled raw centered-score variance:

$$L = L_{\text{pair}} + \lambda_{\text{pres}} L_{\text{preserve}} + \lambda_{\text{reg}}(\|U\|_F^2 + \|V\|_F^2). \tag{5}$$

The hinge asks whether the answer beats its twin. Preservation keeps the relative background scores close to raw, but does not compare the answer with those backgrounds. A readout can therefore improve the paired decision while moving other documents above the answer (Appendix B.1).

Eight preservation weights $\{0, 0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$ and two regularizers $\{10^{-4}, 10^{-3}\}$ give 16 rank-256 trained readouts, or *fits*. Adding raw cosine gives 17 choices per bank. Five seeds for each of six encoders give 30 candidate sets, or *banks*. We train each bank separately by encoder, seed, and task, then freeze its weights before comparing selectors. Raw cosine remains an eligible no-change option. Section 4.4 also uses two separate readouts: a NevIR lens without background preservation (W_{doc}) and a projection built from synthetic affirmative–negative differences (Appendix A).

3.3 What validation sees and how it chooses

A validation view $\mathcal{V} = (Q_{\mathcal{V}}, \mathcal{D}_{\mathcal{V}})$ specifies the queries and document scores visible to a selector. Actual retrieval lists contain only raw cosine's top k documents, for $k = 10, 20$, or 50. An absent answer contributes zero RR@10. A query passes the contrast test if its answer is present and its twin is absent or scores lower; both queries must pass. Only when both documents are present can changing the readout change this comparison. Test contrast always uses the complete labeled pairs.

Controlled lists always contain the answer and twin, plus 0, 1, 2, 8, 32, 128, or all background documents. Backgrounds are either sampled uniformly without replacement or retrieved by raw cosine. Thus, retrieved +8 has ten documents but differs from an actual top-10 list: it guarantees that both labeled documents are present. We also use the full validation corpus and vary the number of pairs over 25, 50, 100, and 225, keeping each pair's two queries together.

Let $\widehat{A}_{\mathcal{V}}(h)$ be the observed pairwise accuracy and $\widehat{R}_{\mathcal{V}}(h) = R_{Q_{\mathcal{V}}, \mathcal{D}_{\mathcal{V}}}(h)$ be the view's MRR@10. Each rule chooses one readout for all test queries:

$$\widehat{h}_{\text{contrast}} = \arg\max_{h \in \mathcal{H}} \widehat{A}_{\mathcal{V}}(h), \tag{6}$$

$$\widehat{h}_{\text{ranking}} = \arg\max_{h \in \mathcal{H}} \widehat{R}_{\mathcal{V}}(h), \tag{7}$$

$$\widehat{h}_{\text{constrained}} = \arg\max_{h \in \mathcal{H}:\, \widehat{R}_{\mathcal{V}}(h) \geq \widehat{R}_{\mathcal{V}}(h_0)} \widehat{A}_{\mathcal{V}}(h). \tag{8}$$

Contrast-only maximizes accuracy; *ranking-first* maximizes MRR. *Ranking-constrained contrast*, or *constrained contrast* for short, discards candidates below raw's validation MRR, then maximizes accuracy among the rest. Ranking is thus an objective for one rule and a filter for the other. This zero-loss filter concerns validation eligibility, not the test-loss tolerance defined below. Ties favor raw, then larger preservation weights, then bank order.

We compute each score matrix once and restrict the entries visible to each selector. At a fixed query count, document count measures how many score entries it can inspect. This is a comparison of validation evidence, not an end-to-end runtime experiment (Appendix B.9).

3.4 Selection with relevance grades

NFCorpus has 3,633 documents with relevance grades 1 and 2 (Boteva et al., 2016; Thakur et al., 2021). *Grade preference* measures grade-2-over-grade-1 decisions, averaged within queries and then across queries with both grades. Ranking uses nDCG@10 over all queries, with gain $2^g - 1$ for grade g and zero for unjudged documents. These metrics replace accuracy and MRR in Equations 6–8.

Training judgments contain only grade 1. Each seed therefore splits the 324 development queries into 162 fitting and 162 selection queries; only 45–54 fitting queries have both grades. We train the same 16-lens grid with a grade-preference hinge. The test set has 323 queries, including 106 with both grades. Seeds vary the split and fitting. Selection uses 40, 80, or all 162 queries with raw lists or the full corpus. Missing-document comparisons follow the rule above, retaining the full grade-pair denominator. List nDCG uses the ideal denominator from all judgments. Appendix D details sampling.

3.5 Available gains and selected outcomes

For a readout h, test contrast gain is $G(h) = A_{\text{test}}(h) - A_{\text{test}}(h_0)$ and full-corpus ranking change is $\Delta R(h) = R_{\text{test}}(h) - R_{\text{test}}(h_0)$. We distinguish three test-based checks:

Useful alternative: a bank member has $G(h) \geq 0.05$ and $\Delta R(h) \geq 0$.
Joint criterion: the selected fit has $G(\widehat{h}) \geq 0.05$ and $\Delta R(\widehat{h}) \geq -0.01$.
Adverse choice: a non-raw selection has $\Delta R(\widehat{h}) < -0.01$.

The contrast target is $\tau_G = 0.05$ (five percentage points); the test-loss tolerance is $\tau_R = 0.01$. The bank diagnostic permits no observed ranking loss, whereas the joint criterion permits up to 0.01. A non-adverse choice need not meet the contrast target. These are post-selection point-estimate checks, not validation rules or guarantees of preservation. Equation 8 instead requires validation MRR to be at least raw's.

After selection, we retrospectively inspect all candidates' test results to identify useful alternatives; this diagnostic never changes the selected readout and is not a prospective selection rule. For NFCorpus, we report the largest observed grade gain and count adverse choices using its own 0.01 nDCG@10 tolerance. Selected readouts are also evaluated on raw's fixed test lists against the matching raw baseline.

Each setting has 20 trials. Query subsets and uniform backgrounds vary across trials; deterministic views with all queries repeat one choice per bank. We average within banks, then weight the 30 banks equally. Banks share evaluation items; bootstrap units are NevIR pairs or NFCorpus queries. NevIR was used during development, and the constrained rule was fixed after the first BGE-M3 run. Appendix C describes the actual-list refits.

4 RESULTS

4.1 USEFUL READOUTS CAN BE MISSED

Every NevIR bank contains one to four candidates with contrast gain ≥ 0.05 and nonnegative test MRR change (Table 6). These post-selection results establish that useful alternatives were observed within each bank, not merely somewhere in the pooled point cloud of Figure 1.

Table 1: **Actual retrieval lists lead to different readout choices.** Each bank uses the same 17 readouts and all 225 validation pairs. Raw top-k contains only returned documents; test contrast uses complete pairs and test MRR the full corpus. Joint: $G \geq 0.05$, $\Delta R \geq -0.01$; adverse: $\Delta R < -0.01$. Counts and means include every choice in all 30 banks.

Validation view	Selection rule	Joint criterion	Adverse choices	Δ pair accuracy	Δ corpus MRR@10
Raw top-10	Contrast-only	0/30	30/30	+0.195	−0.184
Raw top-10	Ranking-first	26/30	0/30	+0.099	+0.006
Raw top-10	Ranking-constrained contrast	22/30	8/30	+0.130	−0.008
Raw top-50	Ranking-first	23/30	0/30	+0.092	+0.006
Raw top-50	Ranking-constrained contrast	23/30	4/30	+0.120	−0.002
Full corpus	Ranking-first	23/30	0/30	+0.092	+0.006
Full corpus	Ranking-constrained contrast	26/30	1/30	+0.117	+0.002

On raw top-10 validation, contrast-only gains 0.195 in mean test accuracy but loses 0.184 corpus MRR, with 30 adverse choices. Ranking-first gains 0.099 accuracy with none; constrained contrast gains 0.130 with eight. Ranking-first therefore retains less contrast gain while avoiding the large ranking losses. All three rules select trained readouts in every bank in this setting, so ranking-first's 0/30 adverse-choice count is not explained by defaulting to raw cosine (Table 1).

The raw validation lists contain the answer for 0.878–0.938 of queries and retrieve both documents for both queries in 0.827–0.907 of pairs. They therefore show most labeled contrasts together with documents that raw ranks highly. Ranking-first meets the joint criterion in 26 banks. The remaining four stay within the loss tolerance but gain less than 0.05 accuracy. Gains vary by encoder: mean accuracy increases range from 0.019 for multilingual E5 to 0.156 for Qwen3-4B (Appendix C).

4.2 COMPETITOR IDENTITY MATTERS BEYOND LIST LENGTH

We keep both labeled documents visible and vary their competitors. With all 225 pairs, ranking-first makes adverse choices in 31.3% of trials with eight uniform backgrounds but none of the 30 banks with eight retrieved backgrounds. Mean contrast gain is $+0.134$ with uniform +8 and $+0.103$ with retrieved +8 (Table 2). Competitor identity therefore changes both the ranking-loss frequency and the retained contrast gain while the bank, rule, and list length stay fixed.

Table 2: **Background identity changes selection at the same list length.** All 225 validation pairs are used. "+8" adds eight background documents to the answer and twin; retrieved backgrounds are highest-scoring under raw. Definitions match Table 1. Uniform rows average 20 draws within banks, then weight the 30 banks equally; retrieved rows give one choice per bank. All non-raw rows select trained fits.

Validation view	Selection rule	Joint criterion	Adverse choices	Δ pair accuracy	Δ corpus MRR@10
No change	Raw cosine	0/30	0/30	$+0.000$	$+0.000$
Labeled pair	Contrast-only	0/30	30/30	$+0.194$	-0.183
Uniform +8	Ranking-first	68.7%	31.3%	$+0.134$	-0.004
Uniform +8	Ranking-constrained contrast	13.0%	87.0%	$+0.162$	-0.042
Retrieved +8	Ranking-first	26/30	0/30	$+0.103$	$+0.005$
Retrieved +8	Ranking-constrained contrast	20/30	10/30	$+0.134$	-0.014

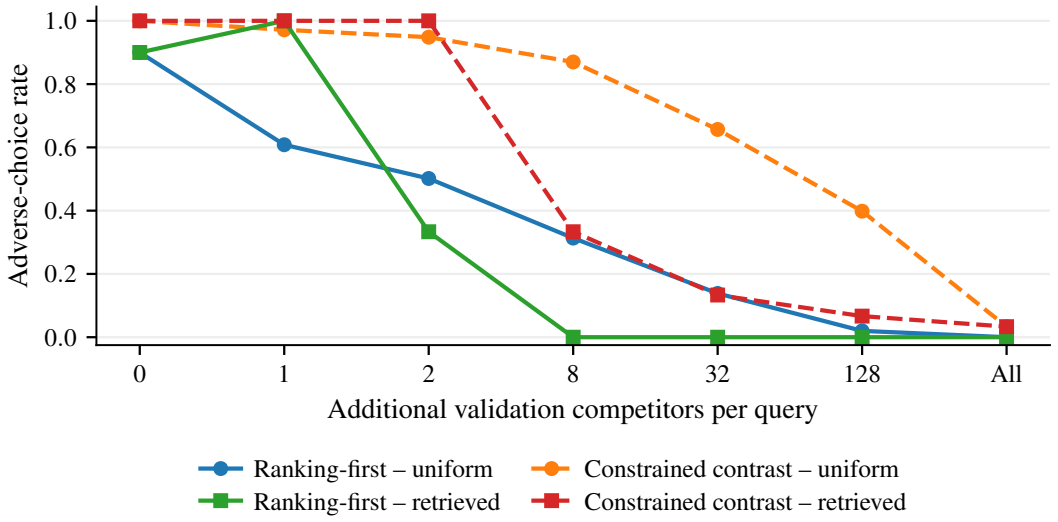


Figure 2: **Competitor source and count affect the frequency of large ranking losses.** Every view includes the answer and twin for all 225 validation pairs. The vertical axis is the fraction of choices with full-test-corpus MRR@10 change below -0.01. Trials are averaged within banks, which receive equal weight. Accuracy gains appear in Appendix B.5.

Retrieved backgrounds are not uniformly better. With one competitor, ranking-first's adverse-choice rate is 100% for retrieved backgrounds and about 61% for uniform ones; with eight, the ordering reverses (Figure 2). Source and count interact: eight is an observed setting, not a universal threshold. At retrieved +8, constrained contrast makes ten adverse choices and retains more contrast gain than ranking-first ($+0.134$ versus $+0.103$).

More examples of a paired contrast do not expose other competitors. Increasing twin-only validation from 25 to 225 pairs raises contrast-only's adverse-choice rate from 79% to 100%. At the same two endpoints, full-corpus ranking-first has adverse-choice rates of 7.2% and 0%, respectively (Table 7).

The added pairs improve the amount of evidence available within each view; they do not change which kinds of document comparisons the view reveals.

4.3 Ranking as an objective or a constraint

The two ranking-based rules see the same MRR values but use them differently. Because raw is eligible, ranking-first's choice is also eligible under the constrained rule. Constrained contrast then chooses the most accurate contrast scorer within that eligible set. It therefore has at least as much validation accuracy and no more validation MRR than ranking-first (Appendix B.3). This ordering need not hold on test data.

Larger validation lists reduce adverse constrained choices from eight at top-10 to four at top-50 and one with the full corpus. Joint-criterion successes increase from 22 to 23 to 26. With full-corpus validation, constrained contrast retains mean contrast gain $+0.117$, compared with $+0.092$ for ranking-first. The latter has 23 joint successes and no adverse choices; constrained contrast has one adverse choice.

Avoiding large losses is not the same as avoiding every loss. With all 225 validation pairs, all twin-only contrast-only choices lose more than 0.05 test MRR. Retrieved +8 ranking-first has two negative changes, each no larger than 0.0016 in magnitude. Appendix B.7 reports the same choices at several loss thresholds.

4.4 The same readout as a reranker or a corpus scorer

We keep the selected functions fixed and change their use. After top-10 validation, ranking-first gains 0.0092 MRR on raw's fixed test top ten. Constrained contrast gains 0.0041 there but loses 0.0077 on the full corpus (Figure 3). These are means over the same 30 selected fits; Table 18 gives encoder means. Contrast-only lowers MRR in both settings.

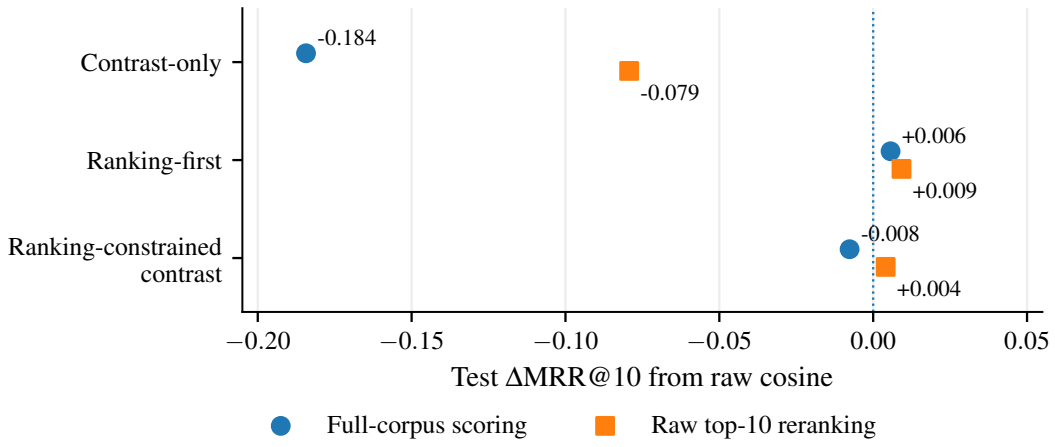


Figure 3: **Average outcomes depend on where the selected readouts are used.** Rules select using raw top-10 validation lists and all 225 pairs. Points average the same 30 choices against each evaluator's raw baseline. These means do not count individual sign reversals. Missing answers contribute zero to fixed-list MRR.

Reranking changes only the order of documents already retrieved. Let $\mathcal{L}_q$ be raw's list, $I_q = \mathbf{1}[d_q^* \in \mathcal{L}_q]$, and $\delta_h(q) = f_{10}(r_h(q; \mathcal{L}_q)) - f_{10}(r_{h_0}(q; \mathcal{L}_q))$, with zero for an absent answer. With equal query weights and positive answer coverage,

$$\Delta\,\mathrm{MRR}_{\mathrm{list}}(\widehat{h}) = \Pr(I_q = 1)\,\mathbb{E}[\delta_{\widehat{h}}(q) \mid I_q = 1]. \tag{9}$$

The fixed-list change is coverage times the conditional mean change, or zero at zero coverage. Reranking cannot recover missing answers. Full-corpus scoring lets every document compete; Equation 9 does not predict that outcome.

To illustrate the role of competitors, write corpus rank as $r_h = 1 + t_h(q) + K_h(q)$: t_h counts a higher-scoring twin and K_h counts all other documents above the answer. In a separate analysis outside the main banks, the canonical projection improves twin order but raises mean K_h from 5.2–14.0 to 167.8–337.7. Its expected MRR gain becomes negative after one uniformly added competitor on five encoders and two on mGTE.

The separately trained W_{doc} lens also shows a sign reversal. On BGE-M3, its expected MRR change is $+0.034$ for the answer–twin pair, $+0.013$ with one background, -0.003 with two, and -0.181 for the full corpus. These examples illustrate a mechanism; they are distinct from the fixed-bank selection experiment. Appendix E averages over background draws with all scores fixed. The sign reversal therefore reflects added competitors, not further changes to the scoring function.

4.5 Small grade-preference gains in the NFCorpus banks

NFCorpus shows a different limitation: the trained candidates offer little additional grade discrimination. Raw grade preference is 0.753–0.807, and the largest test gain among all trained fits is 0.0301. Full-corpus validation selects smaller gains, averaging 0.0011–0.0037 (Table 3; Figure 4).

Table 3: **Small grade-preference gains, different ranking outcomes.** NFCorpus uses 162 selection queries per bank and 323 fixed test queries. Grade preference averages grade-2-over-grade-1 decisions within each of the 106 test queries with both grades, then across those queries. Full-corpus nDCG@10 uses all test queries. Adverse choices lose more than 0.01 nDCG@10; all values summarize 30 banks.

Validation view	Selection rule	Adverse choices	Δgrade preference	Δcorpus nDCG@10
Full corpus	Contrast-only	8/30	$+0.0011$	-0.0089
Full corpus	Ranking-first	0/30	$+0.0037$	$+0.0023$
Full corpus	Ranking-constrained contrast	2/30	$+0.0033$	$+0.0011$
Raw top-10	Contrast-only	5/30	$+0.0007$	-0.0066
Raw top-10	Ranking-first	2/30	$+0.0031$	-0.0016
Raw top-10	Ranking-constrained contrast	3/30	$+0.0006$	-0.0013

With all 162 selection queries and the full corpus, contrast-only makes eight adverse choices, ranking-first none, and constrained contrast two. All choose trained fits. Ranking-first's mean grade gain is $+0.0037$ and nDCG change $+0.0023$. Across encoders, the five-seed mean grade-preference change of the full-corpus ranking-first choices ranges from -0.0032 for Qwen3-4B to $+0.0143$ for multilingual E5 (Appendix D). These small gains describe the current candidates, not an intrinsic limit of graded relevance.

Actual top-10 lists expose fewer relevance judgments: across the 30 encoder–seed selection halves, qrel recall ranges from 0.119–0.187, and the fraction of returned entries with positive judgments ranges from 0.201–0.306. Unlike NevIR, these fractions count multiple annotated documents per query. Ranking-first makes two adverse choices with top-10 validation and none with the full view. Only 45–54 fitting queries supply both grades (Table 19). Limited fitting data, the candidate family, and the task may all affect the small gains; this comparison does not isolate their contributions.

5 Implications and limitations

Interpret selection against the joint goal. Contrast-only can succeed at maximizing accuracy while failing the joint contrast-and-ranking criterion. The fixed banks reveal whether a qualifying alternative was observed. The contribution is to distinguish a missed available gain from limited gains in the candidate bank, rather than to identify the best function the embeddings could support.

Both comparisons and objectives matter. Extra twin-only pairs leave answer–background comparisons unseen. Exposing competitors does not itself prioritize ranking: constrained contrast still chooses for accuracy after filtering. Retrieved +8 and full-corpus ranking-first avoid losses above

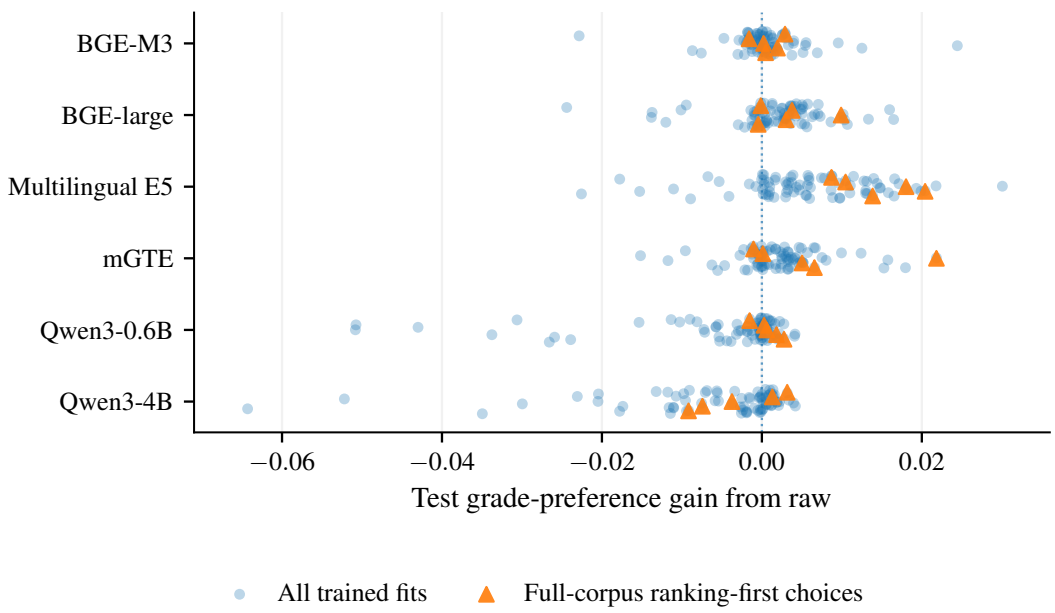


Figure 4: **The trained NFCorpus candidates offer small grade-preference gains.** Each row contains 80 fits for one encoder across five seeds; larger marks show the five full-corpus ranking-first choices. The horizontal axis gives test gain over raw cosine. Vertical offsets separate overlapping points and have no quantitative meaning. Ranking outcomes appear in Table 3.

0.01 with all 225 NevIR pairs; smaller samples do not show the same result. These observations support evaluating the validation view and selection rule together, without a general guarantee of preservation.

Scope of the evidence. NevIR MRR follows designated answers; it does not establish every competitor's relevance. The 30 banks share test items. Useful-candidate diagnoses are post-selection point estimates, not confidence guarantees. Expansion analyses use separate readouts, and NevIR–CondaQA transfer shares a passage source. These distinctions separate the reported observations from independent replication and stronger relevance or uncertainty claims.

Practical use. For prospective use, define the contrast gain and ranking loss that matter for the task. Keep raw eligible and compare selectors on the same view. Select on validation, then evaluate only the validation-selected readout on test against the matching baseline. The all-candidate test inspection used in our analysis is retrospective diagnosis, not a selection step. Report both gains and losses, retain missing answers in list averages, and distinguish fixed-list reranking from full-corpus scoring.

6 Conclusion

Within each selection comparison, holding the candidate bank fixed separates gains available in that bank from gains recovered by validation. On NFCorpus, by contrast, the trained candidate banks offer only small grade-preference gains, leaving little for validation to recover. Competitor identity matters at a fixed list length, and more labeled pairs do not replace missing background comparisons. NFCorpus shows why available gains also require inspection. Finally, reranking and corpus scoring can give different average outcomes for the same selected set. A useful evaluation must examine the candidates, validation choice, and intended retrieval setting.

AI use statement

Generative AI tools were used for manuscript wording and presentation review, and for organizing and documenting reproducibility code and research artifacts. The authors carried out and verified all experiments, analyses, scientific interpretations, and conclusions.

Reproducibility statement

Appendices A–D describe data, training, and selection; Appendix K lists the environment and artifacts. The supplement contains code, result ledgers, and table and figure generators. Its CPU check reproduces the canonical projection analysis. Fixed-bank selection results have separate ledgers.

References

Mohammed Ali, Abdelrahman Abdallah, and Adam Jatowt. EXCISE: Query-side exclusion for late-interaction retrieval. *arXiv preprint arXiv:2608.05497*, 2026.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

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

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 (ECIR)*, pp. 716–722, 2016. doi: 10.1007/978-3-319-30671-1_58.

Hongliu Cao. Semantic adapter for universal text embeddings: Diagnosing and mitigating negation blindness to enhance universality. In *ECAI 2025: 28th European Conference on Artificial Intelligence*, pp. 4137–4144, 2025. doi: 10.3233/FAIA251305.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://jmlr.org/papers/v23/20-1335.html.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Seraphina Goldfarb-Tarrant, Pedro Rodriguez, Jane Dwivedi-Yu, and Patrick Lewis. MultiContrievers: Analysis of dense retrieval representations. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2024. URL https://aclanthology.org/2024.blackboxnlp-1.8/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. CLIP behaves like a bag-of-words model cross-modally but not uni-modally. *arXiv preprint arXiv:2502.03566*, 2025.

Walid Krichene and Steffen Rendle. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1748–1757, 2020. doi: 10.1145/3394486.3403226.

Taegyeong Lee, Jiwon Park, Seunghyun Hwang, and JooYoung Jang. DEO: Training-free direct embedding optimization for negation-aware retrieval. *arXiv preprint arXiv:2603.09185*, 2026.

Daryl Lim and Gert Lanckriet. Efficient learning of Mahalanobis metrics for ranking. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1980–1988, 2014.

Markus Ojala and Gemma C Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863, 2010.

Radoslav Ralev, Aditeya Baral, Iliya Zhechev, Jen Agarwal, and Srijith Rajamohan. Training for compositional sensitivity reduces dense retrieval generalization. *arXiv preprint arXiv:2604.16351*, 2026.

Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. What are you token about? Dense retrieval as distributions over the vocabulary. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. URL https://aclanthology.org/2023.acl-long.140/.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020. doi: 10.18653/v1/2020.acl-main.647.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021. doi: 10.18653/v1/2021.eacl-main.295.

Abhilasha Ravichander, Matt Gardner, and Ana Marasović. CondaQA: A contrastive reading comprehension dataset for reasoning about negation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8729–8755, 2022. doi: 10.18653/v1/2022.emnlp-main.598.

Benjamin Reichman and Larry Heck. Dense passage retrieval: Is it retrieving? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13540–13553, 2024. URL https://aclanthology.org/2024.findings-emnlp.791/.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2021.

Coen van den Elsen, Francien Barkhof, Thijmen Nijdam, Simon Lupart, and Mohammad Aliannejadi. Reproducing NevIR: Negation in neural information retrieval. *arXiv preprint arXiv:2502.13506*, 2025.

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

Orion Weller, Dawn Lawrie, and Benjamin Van Durme. NevIR: Negation in neural information retrieval. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2274–2287, 2024. doi: 10.18653/v1/2024.eacl-long.139.

Xingkun Yin, Xuebin Tang, and Hongyang Du. CoDeR: Local constraint-compatible retrieval beyond semantic similarity. *arXiv preprint arXiv:2606.13204*, 2026.

Wenhao Zhang, Mengqi Zhang, Shiguang Wu, Jiahuan Pei, Zhaochun Ren, Maarten de Rijke, Zhumin Chen, and Pengjie Ren. Excluir: Exclusionary neural information retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13295–13303, 2025a. doi: 10.1609/aaai.v39i12.33451.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1393–1412, 2024. doi: 10.18653/v1/2024.emnlp-industry.103.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025b.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1298–1308, 2019. doi: 10.18653/v1/N19-1131.

GUIDE TO THE EVIDENCE

Three questions organize the appendices. **Selection:** which validation rules choose useful trained candidates? Appendices B–D support Tables 1–3 with bank, retrieval-list, and graded-relevance details. **Expansion:** why can a contrast gain coexist with ranking loss? Appendix E adds competitors under a separate, fixed readout. Selection changes the chosen function; expansion changes the documents competing under one function.

Recovery and scope: Appendices F–I study local reranking, information recoverable from stored vectors, controls, and transfer costs. These studies use their stated readouts and fitting protocols; they are not additional repetitions of the main selection experiment. Appendix A gives data and implementation details, J audits shared sources, and K describes computation and reproducibility.

A DATA AND IMPLEMENTATION

A.1 DATASETS AND MODEL INPUTS

NevIR contains $948/225/1{,}383$ training, validation, and test document pairs, each with two queries whose relevant documents reverse. The test corpus has $2{,}081$ unique documents; readout selection uses a validation corpus of $1{,}761$ unique training and validation documents. The Jina negation resource supplies public training and test triplets. For W_{sent} training, we sample $10{,}000$ valid entailment–contradiction triplets from SNLI train.

NegConstraint has $3{,}946$ documents and 366 queries. ExcluIR (Zhang et al., 2025a) supplies 3,452 exclusionary queries, each paired with a satisfying and a violating document. For the fixed-readout expansion analysis in Appendix E, the candidate corpus is the union of all satisfying and violating documents, and every readout is applied unchanged. The first post-freeze CondaQA check uses 585 train-and-validation groups, followed by chain and transfer evaluations on 688 test groups. NevIR was built from CondaQA passages; their overlap is measured in Appendix J.1 (Table 52). The document-side antonym holdout has 144 near-duplicate pairs from 12 subjects and 12 antonym families, generated independently of fitting and selection data. The archive also contains SNLI, Korean, context-length, and checkpoint diagnostics, separate from the main bank-training and selection runs.

All embeddings are normalized before scoring. Model-specific wrappers follow the released implementations. In particular, multilingual E5 uses the `query:` prefix for queries and sentence-level training items and the `passage:` prefix for NevIR documents. BEIR and NegConstraint use a common 512-token truncation within the released evaluation pipeline. All before–after changes use the same input pipeline and its own raw baseline. Table 4 shows which resources are used for fitting, selection, and evaluation.

Table 4: Use of each resource in fitting, selection, and evaluation.

Resource	Fitting and selection	Evaluation role	Metric
NevIR	lens training on train; readout choice on validation	contrast, candidate expansion, and selection outcomes on test; used during development	pair acc. / MRR
Jina + SNLI	W_{sent} training; projection rank and scale selected on Jina train	external training and selection	triplet AUC
Synthetic pairs	projection-basis and external-span construction	independent geometry and transfer controls	energy / acc.
CondaQA	projection frozen before first check; later lenses use its train and validation splits	585 train+validation groups initially; 688 test groups later	pair acc.
Antonym families	none	144 adversarial near-duplicate pairs	pair acc.
NFCorpus	dev split into fitting and selection halves	fixed test queries; graded preference and ranking	grade pref. / nDCG
NegConstraint + other BEIR collections	none	contract and local-policy evaluations	nDCG@10
ExcluIR	none	fixed-readout exclusion and candidate-expansion evaluation	expected MRR@10
PAWS	task-specific probe/lens fitting and validation	argument-order probe and retrieval evaluation	pair acc. / AUC

A.2 Accessibility probes

The cue detector is a fixed regular expression over explicit English negation markers and contractions. Sentence probes use a fixed shuffled split. TF–IDF is fitted on the same split; cue deletion occurs on raw text before vectorization. The document probe uses only pairs with unequal cue counts. Writing ϕ for an ordered document-embedding difference and y for its binary cue-count label, each original pair contributes (ϕ, y) and $(-\phi, 1-y)$; the two rows remain in the same train/test group and bootstrap cluster. Mirroring gives exact class balance; the original document pair remains the independent unit. Single shuffled-label fits provide a preliminary check, and the repeated permutation tests below supply the reference distributions.

A.3 Learned lenses

Both W_{sent} and W_{doc} apply $T_{U,V}(z) = \text{norm}(z + UV^\top z)$ to the query and document vectors. We initialize $U, V \in \mathbb{R}^{D\times r}$ from $\mathcal{N}(0, 0.01^2)$ and train them with Adam at learning rate 10^{-3} and penalty $\lambda(\|U\|_F^2 + \|V\|_F^2)$. Training runs for the full fixed number of epochs. Validation chooses (r, λ) from $\{(64, 10^{-4}), (256, 10^{-4}), (256, 10^{-3})\}$.

The document-trained lens, W_{doc}, minimizes $\max(0, \mu - (s_{11} - s_{12})) + \max(0, \mu - (s_{22} - s_{21}))$ on NevIR train, with the same hinge margin $\mu = 0.05$. Here s_{ij} is the transformed cosine between query i and document j. Training uses batches of 128 pairs for 40 epochs (seed 0; three seeds in Appendix G.1). BGE-M3 selects $(256, 10^{-3})$.

The sentence-trained lens, W_{sent}, uses InfoNCE at temperature 0.05 on (anchor, positive, negative) triplets from Jina and 10,000 SNLI triplets. It uses batches of 256 for 3 epochs, followed by selection on NevIR validation. The three BGE-M3 runs score 0.321, 0.328, and 0.317; the repeated-fit analysis is in Appendix G.1.

Which fit each table reports. The recovery study selects rank and regularization; the later corpus and transfer studies refit at rank 256. The standalone preservation study and the selection study also train separate banks. Each table pairs its fit with the raw baseline from the same run, identified by the settings and run ID in the result records. Changes in precision, pooling, and input wrappers account for third-decimal differences between some raw baselines. Comparisons use the matched before–after values within each table.

The score after a shared transform. Substituting $M = I + UV^\top$ from Equation 4 and writing $H = M^\top M$, the shared-transform score is

$$s_h(q, d) = \frac{z(q)^\top H z(d)}{\sqrt{z(q)^\top H z(q)}\sqrt{z(d)^\top H z(d)}}. \tag{10}$$

Normalizing the transformed vectors gives the denominator. The learned map changes both coordinate combinations and their norms. An orthogonal M gives $H = I$ and leaves cosine unchanged; a zero residual update also gives raw cosine. The general lens in Equation 10 is a normalized metric score, with both its numerator and denominator set by H.

A.4 Canonical projection

The construction set contains 144 English propositions transformed with explicit *not*, *no*, *never*, and *cannot* forms. Two negative forms are retained per proposition and affirmative paraphrases are excluded, yielding 288 difference vectors. We compute an SVD separately for each encoder, search $k \in \{1, 2\}$ and $\alpha \in \{0.5, 1, 2, 4, 8\}$ on Jina training-triplet AUC, and freeze the selected projection. BGE-M3 selects $(k, \alpha) = (2, 4)$ and mGTE selects $(1, 8)$. The manifest stores the full hash of the projection matrix $P^\top P$ as well as a short basis identifier. For an orthonormal row basis P, the transformation is

$$T_{P,\alpha}(z) = \mathrm{norm}(z + \alpha P^\top P z). \tag{11}$$

Before normalization, the selected component is scaled by $1 + \alpha$; its orthogonal component stays unchanged. We use this external projection for recovery and candidate expansion. Selection banks contain the learned lenses in Section 3.2.

A.5 Metrics and inference

NevIR pairwise success requires both directional queries to rank the correct member of their document pair. NegConstraint and BEIR use nDCG@10 with the same implementation before and after intervention. Pair-bootstrap intervals resample original NevIR document pairs; query-bootstrap intervals resample complete retrieval queries. On the NevIR corpus contract the resampled unit is the document pair with both of its queries, because the two queries of a pair are dependent; on the other collections it is the query. Paired NevIR comparisons use two-sided exact McNemar tests. The sweeps describe performance across six values of α; their intervals are pointwise and unadjusted for multiple comparisons.

B Readout selection: protocol and full results

The main selection experiments compare rules that receive the same trained alternatives. This appendix specifies the preservation loss, the shared-bank protocol, the available and selected gains, and the full validation sweeps. The last subsection reports the separately trained preservation study. Actual-list views and graded-relevance selection have their own protocols in Appendices C and D.

B.1 What the preservation loss computes

The main bank uses the `preserve` objective in Equation 5: the answer–twin hinge, background-score preservation, and factor regularization. For a minibatch of B document pairs, draw $J = 16$ training-document indices for each pair. The two queries of pair i use the same background draw $(b_{i1}, \ldots, b_{iJ})$. The implementation samples with replacement and resamples entries that match either member of that pair. For query role $a \in \{1, 2\}$, define

$$c^{(a)}_{h,ij} = s_h(q_{ia}, b_{ij}) - \frac{1}{J}\sum_{u=1}^{J} s_h(q_{ia}, b_{iu}). \tag{12}$$

This subtraction removes an overall shift in one query's background scores. It leaves their differences unchanged. The loss in the implementation is

$$L_{\text{preserve}} = \frac{1}{2}\sum_{a=1}^{2} \frac{\frac{1}{BJ}\sum_{i=1}^{B}\sum_{j=1}^{J}(c_{h,ij}^{(a)} - c_{h_0,ij}^{(a)})^2}{\max\left(\frac{1}{BJ}\sum_{i=1}^{B}\sum_{j=1}^{J}(c_{h_0,ij}^{(a)})^2, 10^{-8}\right)}. \tag{13}$$

The numerator averages squared score changes over the minibatch and its backgrounds. Each query role has one denominator: the raw centered-score variance pooled over that minibatch. The floor 10^{-8} prevents division by zero, and $1/2$ averages the two roles. This penalty uses the background scores, while the contrast hinge in Equation 5 uses the answer–twin labels.

B.2 Shared fitted alternatives and trial units

The selection experiment refits the 16-member preservation grid independently for each encoder and seed, and adds raw cosine. The bank has the same hyperparameter design as the standalone preservation experiment, but is fitted separately. Within one bank all selectors see identical fits, query subsets, and candidate documents. In the constructed views reported in this appendix, each query's designated answer and twin are always included. As defined in Section 3, the pooled-corpus outcome tracks that fixed target. Uniform backgrounds are drawn without replacement; raw-top backgrounds are the highest-scoring other documents under raw cosine. Thus raw-top includes both labeled documents by construction. Appendix C evaluates the unmodified top-k output.

NevIR was inspected during development. The constrained rule was fixed after the initial BGE-M3 development run and then applied to the remaining encoder conditions. These conditions share the evaluation items; their repeated choices test stability across fits within this source. Raw cosine remains a choice in every bank and wins ties. It provides a no-change option. The outcome measures below distinguish that choice from adopting a trained readout.

Each setting has 20 trials. Smaller pair subsets and uniform backgrounds can change between trials. At all 225 pairs, however, retrieved and full-corpus settings repeat the same observations and choices. These settings yield 30 distinct bank outcomes. Tables average trials within each bank and then weight the banks equally.

B.3 How a selected readout is scored

The ranking-first fit is feasible for the constrained selector because raw is in the bank. Equations 7–8 therefore imply, on the same validation view,

$$\widehat{A}_{\mathcal{V}}(\widehat{h}_{\text{constrained}}) \geq \widehat{A}_{\mathcal{V}}(\widehat{h}_{\text{ranking}}), \qquad \widehat{R}_{\mathcal{V}}(\widehat{h}_{\text{ranking}}) \geq \widehat{R}_{\mathcal{V}}(\widehat{h}_{\text{constrained}}). \tag{14}$$

These inequalities describe the choices on their validation view. The test comparison evaluates the resulting readouts on the held-out queries.

One trial chooses one readout $\widehat{h}$ for all test queries. Write A_{test} and R_{test} for pairwise accuracy and full-corpus MRR@10 on the test set. Its contrast gain and ranking change are

$$G(\widehat{h}) = A_{\text{test}}(\widehat{h}) - A_{\text{test}}(h_0), \qquad \Delta R(\widehat{h}) = R_{\text{test}}(\widehat{h}) - R_{\text{test}}(h_0). \tag{15}$$

We record $\mathbf{1}[\widehat{h} \neq h_0]$ for adoption and $\mathbf{1}[\widehat{h} \neq h_0 \text{ and } \Delta R(\widehat{h}) < -\tau_R]$ for an adverse choice. Thus choosing raw gives no gain and no adverse adoption. The joint-criterion count asks whether $G(\widehat{h}) \geq \tau_G$ and $\Delta R(\widehat{h}) \geq -\tau_R$. Both thresholds classify the completed test outcome. Ranking-first maximizes validation MRR, and constrained selection compares validation MRR directly with raw.

To measure a missed opportunity after selection, the test oracle finds

$$G^* = \max_{h \in \mathcal{H}:\, \Delta R(h) \geq -\tau_R} G(h).$$

The recorded regret is $G^* - G(\widehat{h})$ when the chosen readout is within tolerance and $G^* - G(h_0) = G^*$ otherwise. We compute this oracle after validation has selected the readout. It supplies a test-based reference for missed gain. Table means include every choice, both within and beyond the loss tolerance.

Table 5: **Following a readout choice from training to use.** The protocol distinguishes candidate training, validation selection, evaluation, and post-selection inspection.

Stage	What is held fixed or changed	What to record
Fit	Train the alternatives, add raw, and then freeze every candidate's weights.	The candidate grid and each bank's run identifier.
Choose	Give competing rules the same validation view. Change the view separately.	The visible queries and documents, selection objective, and chosen fit.
Evaluate	Use the validation-selected fit on the target corpus or the first-stage lists; do not alter that selection based on test outcomes.	Contrast gain, ranking change, loss size, and answer coverage.
Inspect	After evaluating the selected fit, retrospectively compare the recorded test outcomes of the other fits for diagnosis only; do not feed this inspection back into selection.	Which gains were available and which validation obtained.

B.4 Available alternatives within the same banks

After evaluating the choices, we inspect the recorded test metrics of all 17 readouts in each bank. A fit meets the bank diagnostic when its contrast gain is at least 0.05 and its corpus MRR change is nonnegative. Table 6 reports the number of such fits and their gains. Every bank contains at least one, so the contrast-only failures in Section 4.1 can be compared with a recorded alternative inside that same bank.

The bank diagnostic requires $G(h) \geq 0.05$ and $\Delta R(h) \geq 0$; the selected-outcome joint criterion requires $G(\widehat{h}) \geq 0.05$ and $\Delta R(\widehat{h}) \geq -0.01$ (Section 3.5). Both summarize test point estimates after selection. The analysis script identifies each qualifying fit and checks its contrast and corpus metrics in the constructed-view and actual-list records. Figure 1 shows all 480 trained alternatives and the three sets of actual-top-10 choices at their recorded test coordinates. The 30 choices per rule share evaluation items; overlapping points can represent several bank members.

Table 6: **Useful alternatives recorded within the NevIR selection banks.** After selection, we inspect each bank for fits with test contrast gain at least 0.05 and full-test-corpus MRR change at least zero. These point-estimate criteria summarize the bank after validation selection. The final column ranges over the fits meeting both conditions in the encoder's five banks. This diagnostic requires nonnegative MRR change; the main joint criterion permits a decrease of 0.01. The two recorded test metrics match between the constructed-view and actual-list banks.

Encoder	Banks with such a fit	Fits per bank	Contrast gains of those fits
BGE-M3	5/5	2–3	$+0.054$ to $+0.124$
BGE-large	5/5	2–2	$+0.082$ to $+0.093$
Multilingual E5	5/5	1–1	$+0.062$ to $+0.068$
mGTE	5/5	4–4	$+0.063$ to $+0.147$
Qwen3-0.6B	5/5	2–4	$+0.051$ to $+0.142$
Qwen3-4B	5/5	3–3	$+0.069$ to $+0.194$

Interpreting the useful-candidate count. Table 6 counts alternatives within each bank, not choices made by a validation rule. The contrast-only result can therefore fail the joint goal even though another trained candidate meets it on the observed test set. This is a diagnostic comparison after selection. Searching the candidates' test scores does not provide a rule for selecting on future data, and nonnegative observed MRR change is not a statistical guarantee of preservation.

B.5 SELECTION CURVES AND VALIDATION-PAIR COUNTS

Table 7 varies the number of validation pairs. Figure 5 shows the contrast gains for the same choices whose losses appear in the main text. Both summarize the choices made in the original validation trials.

Table 7: **More contrast pairs leave background comparisons unobserved.** Each cell gives the fraction of choices beyond the 0.01 loss tolerance / mean test accuracy gain. Both queries stay with their document pair. Trials are averaged within banks before equally weighting the 30 banks. At 225 pairs, retrieved and full-corpus repetitions coincide. Candidate counts measure the scoring information shown to a selector.

Pairs	Contrast-only Labeled pair	Ranking-first Retrieved + 8	Ranking-first Full corpus	Constrained contrast Full corpus
25	79.0% / +0.171	28.3% / +0.104	7.2% / +0.083	17.3% / +0.094
50	89.5% / +0.184	17.8% / +0.108	7.5% / +0.095	27.2% / +0.113
100	95.5% / +0.191	4.7% / +0.097	0.5% / +0.093	17.8% / +0.116
225	100.0% / +0.194	0.0% / +0.103	0.0% / +0.092	3.3% / +0.117

Interpreting the sample-size sweep. In Table 7, more pairs add examples to a fixed kind of validation view. Twin-only views still omit answer–background comparisons; full-corpus views expose them. Twenty repetitions with all pairs do not create twenty independent choices when the documents and rule are deterministic. They repeat the same bank outcome. Averaging across banks describes stability across encoders and fitted seeds on shared evaluation items, rather than replication across independent datasets.

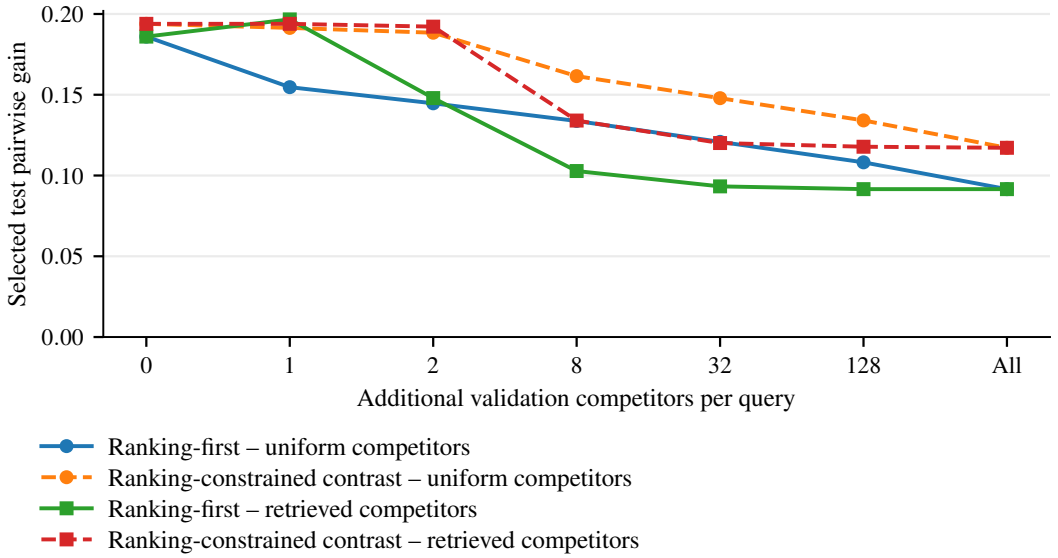


Figure 5: **Contrast gains from the choices in Figure 2.** Each point includes all choices, including those beyond the corpus-loss tolerance. Trials are averaged within banks and then banks receive equal weight. Raw cosine has zero change.

B.6 SELECTION OUTCOMES ACROSS VALIDATION SETTINGS

Tables 8 and 9 compare background sources at each candidate count. Table 10 changes the number of labeled validation pairs, while Table 11 breaks out the highlighted settings by encoder. Each table includes the gains of choices that exceed the loss tolerance, as well as those that stay within it. Rule names follow Section 3.3.

Table 8: Selection with all 225 validation pairs and uniformly sampled backgrounds. The adverse rate counts test-corpus MRR@10 losses greater than 0.01. Gains and regret include all choices. We average trials within banks, then weight the 30 banks equally. The last block uses the auxiliary count plug-in described in Appendix B.

Extra documents	Adverse rate	Δpair accuracy	Δcorpus MRR	Oracle regret
Contrast-only				
0	1.00	+0.194	−0.183	0.126
1	1.00	+0.194	−0.183	0.126
2	1.00	+0.194	−0.183	0.126
8	1.00	+0.194	−0.183	0.126
32	1.00	+0.194	−0.183	0.126
128	1.00	+0.194	−0.183	0.126
All	1.00	+0.194	−0.183	0.126
Ranking-first				
0	0.90	+0.186	−0.122	0.113
1	0.61	+0.155	−0.035	0.069
2	0.50	+0.145	−0.016	0.056
8	0.31	+0.134	−0.004	0.037
32	0.14	+0.121	+0.001	0.027
128	0.02	+0.108	+0.006	0.021
All	0.00	+0.092	+0.006	0.034
Ranking-constrained contrast				
0	1.00	+0.194	−0.183	0.126
1	0.97	+0.191	−0.123	0.123
2	0.95	+0.188	−0.117	0.120
8	0.87	+0.162	−0.042	0.107
32	0.66	+0.148	−0.025	0.082
128	0.40	+0.134	−0.010	0.051
All	0.03	+0.117	+0.002	0.014
Constrained contrast (count plug-in)				
1	0.94	+0.184	−0.109	0.118
2	0.91	+0.165	−0.056	0.113
8	0.72	+0.152	−0.029	0.088
32	0.44	+0.137	−0.013	0.058
128	0.31	+0.130	−0.006	0.041

Table 11: Encoder-level outcomes for two observed selection settings. Each row averages five banks using all 225 validation pairs. Adverse choices are test-corpus losses greater than 0.01 MRR@10. Retrieved validation includes the relevant document, twin, and eight background competitors.

	Ranking-first Retrieved + 8			Ranking-constrained contrast Full corpus		
Encoder	Adverse	Δpair	ΔMRR	Adverse	Δpair	ΔMRR
BGE-M3	0/5	+0.097	+0.003	1/5	+0.124	−0.008
BGE-large	0/5	+0.089	+0.006	0/5	+0.087	+0.009
Multilingual E5	0/5	+0.019	+0.003	0/5	+0.036	+0.007
mGTE	0/5	+0.127	+0.011	0/5	+0.138	+0.010
Qwen3-0.6B	0/5	+0.120	+0.001	0/5	+0.120	+0.001
Qwen3-4B	0/5	+0.165	+0.008	0/5	+0.197	−0.005

B.7 Reading the test-loss tolerance

The main adverse-choice count uses $\Delta R < -\tau_R = -0.01$. To show which part of the result depends on that cutoff, we retain each selected readout and recount the event $\Delta R < -\delta$ at $\delta \in \{0, 0.005, 0.01, 0.02, 0.05\}$. This post-selection analysis reclassifies the recorded outcomes of the original validation choices.

Table 9: Full selection results with all 225 validation pairs and backgrounds retrieved by raw cosine. The relevant document and twin are always included. At this validation size the repeated trials are identical within each bank; rates summarize 30 bank outcomes. Metric definitions and weighting match Table 8.

Extra documents	Adverse rate	Δpair accuracy	Δcorpus MRR	Oracle regret
Contrast-only				
1	1.00	+0.194	−0.183	0.126
2	1.00	+0.194	−0.183	0.126
8	1.00	+0.194	−0.183	0.126
32	1.00	+0.194	−0.183	0.126
128	1.00	+0.194	−0.183	0.126
Ranking-first				
1	1.00	+0.197	−0.176	0.126
2	0.33	+0.148	−0.036	0.042
8	0.00	+0.103	+0.005	0.023
32	0.00	+0.093	+0.006	0.033
128	0.00	+0.092	+0.006	0.034
Ranking-constrained contrast				
1	1.00	+0.194	−0.183	0.126
2	1.00	+0.192	−0.131	0.126
8	0.33	+0.134	−0.014	0.054
32	0.13	+0.120	−0.002	0.025
128	0.07	+0.118	+0.001	0.018

Table 10: Validation sample size for contrast-only and ranking-constrained contrast selection, with uniform backgrounds. Each document pair retains both directional queries. Entries are adverse-choice rates and selected test accuracy gains; trial and bank averaging follow Table 8.

		Contrast-only		Ranking-constrained contrast	
Pairs	Extra documents	Adverse	Δpair	Adverse	Δpair
25	0	0.79	+0.171	0.77	+0.167
25	8	0.83	+0.176	0.60	+0.146
25	All	0.79	+0.171	0.17	+0.094
50	0	0.90	+0.184	0.90	+0.184
50	8	0.90	+0.183	0.68	+0.150
50	All	0.90	+0.184	0.27	+0.113
100	0	0.95	+0.191	0.95	+0.191
100	8	0.93	+0.189	0.74	+0.158
100	All	0.95	+0.191	0.18	+0.116
225	0	1.00	+0.194	1.00	+0.194
225	8	1.00	+0.194	0.87	+0.162
225	All	1.00	+0.194	0.03	+0.117

All 225 validation pairs are used. The labeled-pair, retrieved-background, and full-corpus settings have a single recorded choice per bank: each has modal share one, and the selected fit's stored score matches the cell summary. We trace that choice to the bank's test scores, reproduce the original 0.01 count, and apply the other cutoffs. Uniform-background summaries retain the original 0.01 event rates; the present reclassification uses the fixed choices that can be traced individually.

Table 12: **The same validation choices at different test-loss thresholds.** Each cell counts non-raw choices with $\Delta R_{\text{test}} < -\delta$ among the 30 banks using all 225 validation pairs. The deterministic choices from Table 2 stay fixed as the cutoff changes. Each choice is traced to its stored test result. Table 2 retains the sampled-uniform rates at the original 0.01 cutoff. All banks share evaluation items.

Validation and selection	$\delta = 0$	0.005	0.01	0.02	0.05
Raw cosine	0/30	0/30	0/30	0/30	0/30
Contrast-only	30/30	30/30	30/30	30/30	30/30
Retrieved + 8 Ranking-first	2/30	0/30	0/30	0/30	0/30
Retrieved + 8 Constrained contrast	11/30	10/30	10/30	10/30	5/30
Full corpus Ranking-first	1/30	0/30	0/30	0/30	0/30
Full corpus Constrained contrast	9/30	5/30	1/30	1/30	0/30

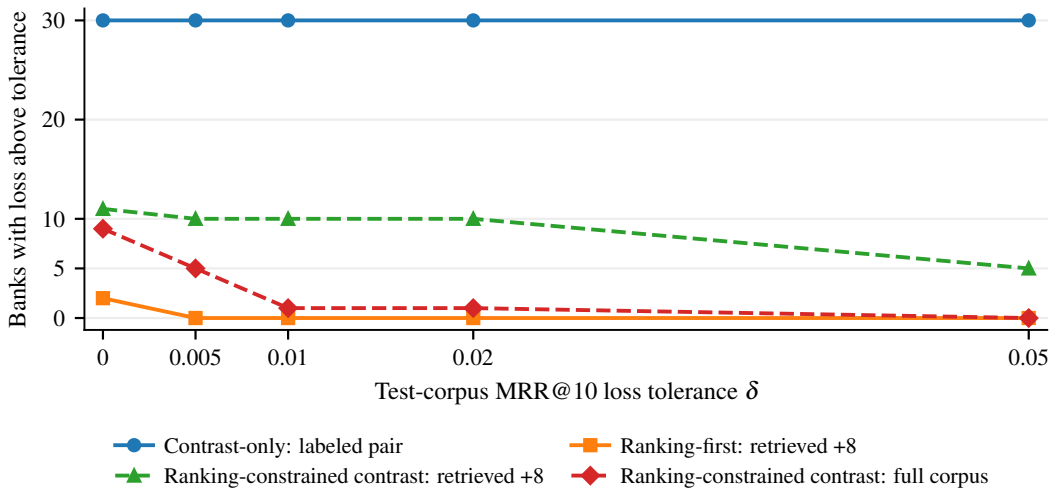


Figure 6: **The recorded choices at different loss cutoffs.** The original validation choices stay fixed while the test-loss cutoff changes. Counts summarize the 30 banks using all validation pairs. The source records, cutoffs, and exact counts appear in Table 12.

All contrast-only choices lose more than 0.05 MRR. Retrieved-background ranking-first has two negative changes, neither above 0.005; the maximum loss is 0.0016. Full-corpus constrained selection has nine negative changes, including five above 0.005 and one above 0.02. This threshold sweep shows how loss magnitudes differ across rules. The script `tools/build_tolerance_sensitivity.py` records each choice and its source-ledger hash.

B.8 Estimating a full-corpus score from sampled counts

For uniform sampling, the auxiliary selectors stored as `R1_corrected` and `R2_corrected` let N be the validation background-pool size and $t_h(q)$ indicate whether the twin scores above the answer. For query q and readout h, they scale the observed overtaker count $X_h(q)$ by N/m and evaluate $f_{10}(1 + t_h(q) + \widehat{K}_h(q))$, where $\widehat{K}_h(q) = (N/m)X_h(q)$ and $f_{10}(r) = \mathbf{1}[r \le 10]/r$. The first step

estimates the number of documents above the answer; the second turns that estimate into reciprocal rank. Reciprocal rank is nonlinear, so an accurate mean count need not give an accurate mean score. This plug-in is distinct from the optimized corrections of Krichene & Rendle (2020).

B.9 Observed information and computation

The simulation computes every validation score matrix, then restricts each selector to the entries in its view. Candidate counts measure the scoring information it reads. Timings cover the selection step after matrix construction. For example, 25 pairs and the complete 1,761-document corpus expose 88,050 query–document scores per fit, whereas 225 pairs with the labeled pair and eight background documents expose 4,500. Choosing raw-top competitors additionally requires the baseline's scores or retrieval results.

Code-to-paper terminology. The immutable result records use `pair_only`, `R1`, and `R2` for contrast-only, ranking-first, and ranking-constrained contrast selection. The `rawtop` key denotes competitors retrieved by raw cosine, and `corrected` denotes the count-extrapolated plug-in just described. These identifiers link the displayed names to the immutable records.

B.10 Background-preserving fits before selection

This standalone experiment tests whether the lens family can provide useful alternatives. It trains its own fits with the objective and full-corpus constrained contrast rule in Section 3. Its recorded success criterion combines pairwise gain of at least 0.05, McNemar $p < 0.05$, test corpus change of at least -0.01, and a 95% interval with a nonnegative upper endpoint. Both zero-crossing and entirely positive intervals pass. This descriptive screen differs from a noninferiority test. The main selection study counts its joint criterion using the two point-estimate thresholds.

Table 14 shows the preservation-weight sweep at the selected regularization, so the reader can compare the chosen fit with its neighbors. A further variant penalizes the relevant document losing ground to a background document. It meets the same criterion on three of six encoders; selected pairwise gains across the six encoders range from $+0.043$ to $+0.223$. Across five training seeds, the criterion holds on 25 of 30 encoder–seed runs: on every seed for BGE-large, mGTE, Qwen3-0.6B and Qwen3-4B; the selected gain ranges over $+0.017$ to $+0.202$ and the corpus change over -0.042 to $+0.012$ across all runs. Within each encoder, the standard deviation of the gain is 0.004 to 0.024. The selected λ is the same on every seed for two of six encoders (Table 15).

Table 15: The preserving-lens procedure across five seeds, which change initialization and background samples. Each seed cell gives preservation weight / test contrast gain on its first line and test corpus change / criterion result on its second. Seed 0 is the run in Table 13. The last column counts seeds meeting the stated criterion.

Model	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4	Holds
BGE-M3	0.03 / +0.127 −0.003 / y	0.03 / +0.119 −0.003 / y	0.1 / +0.148 −0.042 / n	0.1 / +0.151 −0.040 / n	0.03 / +0.126 −0.002 / y	3/5
BGE-large	0.03 / +0.091 +0.005 / y	0.3 / +0.080 +0.010 / y	0.3 / +0.082 +0.009 / y	0.3 / +0.080 +0.009 / y	0.3 / +0.086 +0.012 / y	5/5
multilingual E5	0.3 / +0.020 +0.004 / n	0.1 / +0.063 +0.011 / y	0.3 / +0.017 +0.004 / n	0.1 / +0.065 +0.011 / y	0.3 / +0.021 +0.005 / n	2/5
mGTE	0.01 / +0.134 +0.008 / y	0.01 / +0.143 +0.010 / y	0.01 / +0.139 +0.011 / y	0.01 / +0.133 +0.010 / y	0.01 / +0.132 +0.009 / y	5/5
Qwen3-0.6B	0.3 / +0.127 +0.002 / y	0.3 / +0.121 +0.003 / y	0.3 / +0.119 +0.003 / y	0.3 / +0.119 +0.004 / y	0.3 / +0.117 −0.001 / y	5/5
Qwen3-4B	0.3 / +0.190 +0.002 / y	0.01 / +0.197 −0.007 / y	0.01 / +0.202 −0.008 / y	0.01 / +0.202 −0.007 / y	0.01 / +0.202 −0.008 / y	5/5

Table 13: Background-preserving lenses selected by validation. Each lens uses the same W_{doc} family and training pairs, with a penalty on changes in query-score differences among 16 background documents. Validation chooses the highest pairwise accuracy among fits whose corpus MRR@10 is at least raw. Panel (a) gives reference scores as pairwise accuracy / corpus MRR@10 and the selected contrast gain. Panel (b) gives its corpus change with a pair-clustered 95% interval, and nDCG@10 changes on NegConstraint / SciFact. The plain lens is the $\lambda = 0$ column of Table 14. Coexistence requires pairwise gain at least 0.05 (McNemar $p < 0.05$), corpus change at least -0.01, and a nonnegative interval upper endpoint. Both zero-crossing and positive intervals qualify under this descriptive screen. Appendix B.10 distinguishes it from a noninferiority test.

(a) Reference scores and the selected contrast result

Model	Raw cosine	Canonical	λ	Lens pairwise accuracy (gain)
BGE-M3	0.267 / 0.577	0.408 / 0.282	0.03	0.394 (+0.127)
BGE-large	0.256 / 0.561	0.371 / 0.206	0.03	0.347 (+0.091)
multilingual E5	0.188 / 0.591	0.305 / 0.265	0.3	0.208 (+0.020)
mGTE	0.148 / 0.556	0.372 / 0.393	0.01	0.282 (+0.134)
Qwen3-0.6B	0.153 / 0.528	0.338 / 0.201	0.3	0.281 (+0.127)
Qwen3-4B	0.195 / 0.566	0.376 / 0.264	0.3	0.385 (+0.190)

(b) Corpus rank, other contracts, and criterion

Model	Corpus MRR@10 (change and 95% interval)	NegC. / SciFact	Coexistence
BGE-M3	0.575 (−0.003 [-0.011,+0.006])	−0.006 / −0.027	yes
BGE-large	0.567 (+0.005 [-0.002,+0.013])	−0.001 / −0.013	yes
multilingual E5	0.595 (+0.004 [+0.001,+0.007])	+0.002 / +0.001	no
mGTE	0.564 (+0.008 [+0.001,+0.016])	−0.002 / −0.052	yes
Qwen3-0.6B	0.529 (+0.002 [-0.006,+0.010])	+0.005 / −0.011	yes
Qwen3-4B	0.568 (+0.002 [-0.008,+0.012])	+0.002 / −0.035	yes

Table 14: The preservation frontier. NevIR test pairwise accuracy above test corpus MRR@10 (two lines per cell) of the lens at each preservation weight λ (columns), at the regularization weight of the selected fit (in parentheses); $\lambda = 0$ is the plain W_{doc} lens. Raw cosine and the both-sides projection are in Table 13.

	Preservation weight λ							
Model (λ_{reg})	0	0.01	0.03	0.1	0.3	1	3	10
BGE-M3 (0.001)	0.458 0.396	0.421 0.544	0.394 0.575	0.323 0.580	0.283 0.580	0.270 0.578	0.267 0.577	0.268 0.577
BGE-large (0.001)	0.403 0.439	0.372 0.546	0.347 0.567	0.299 0.572	0.275 0.565	0.261 0.562	0.257 0.561	0.256 0.561
multilingual E5 (0.0001)	0.381 0.132	0.362 0.472	0.316 0.570	0.254 0.601	0.208 0.595	0.192 0.592	0.190 0.591	0.187 0.591
mGTE (0.001)	0.350 0.505	0.282 0.564	0.225 0.568	0.187 0.563	0.162 0.558	0.152 0.557	0.150 0.556	0.149 0.556
Qwen3-0.6B (0.0001)	0.375 0.092	0.387 0.307	0.375 0.384	0.340 0.482	0.281 0.529	0.206 0.535	0.163 0.530	0.156 0.528
Qwen3-4B (0.0001)	0.423 0.110	0.437 0.265	0.435 0.369	0.424 0.484	0.385 0.568	0.275 0.581	0.221 0.570	0.198 0.567

C SELECTION FROM ACTUAL RETRIEVAL LISTS

C.1 KEEPING ONLY THE DOCUMENTS RETURNED BY RAW RETRIEVAL

The earlier retrieved-background design keeps each designated answer and twin and adds raw cosine's highest-scoring other documents. The actual-list study (recorded as E5) instead takes the unmodified raw top-k, for $k \in \{10, 20, 50\}$. Candidate lists are built once from raw scores and reused by every fitted readout. These lists contain exactly the returned documents. The full-corpus view shows all 1,761 validation documents, and both views use the same test corpus and targets.

Each run uses raw cosine plus the same 16 hyperparameter settings as the earlier fixed-bank experiment, with seed-specific fits. The shipped record compares test pairwise gains and test-corpus MRR changes for all 17 candidates against the earlier bank: the maximum absolute difference is zero in every one of the 30 encoder–seed runs. The agreement field checks only these two recorded test metrics. Because the actual-list fits were refit separately, this establishes metric agreement with the earlier bank, not parameter identity; parameter hashes were outside the agreement check. Within the actual-list study, the resulting refitted bank is then held fixed across selectors and validation views. Fitted matrices remain in the working cache, and the compact release supports refitting from cached vectors.

C.2 A CONTRAST THAT CAN BE OBSERVED ON A RETURNED LIST

A list can omit either labeled document. For query q, target d_q^*, twin d_q^{tw}, and raw list $\mathcal{L}_q$, the validation contrast indicator is

$$c_h(q, \mathcal{L}_q) = \mathbf{1}[d_q^* \in \mathcal{L}_q]\, \mathbf{1}[d_q^{\mathrm{tw}} \notin \mathcal{L}_q \ \vee \ s_h(q, d_q^*) > s_h(q, d_q^{\mathrm{tw}})]. \tag{16}$$

Both queries must satisfy this condition for a pair to count as recovered. If the target is absent, the outcome is zero for every candidate. If only the twin is absent, the contrast outcome is one for every candidate; that query alone does not distinguish the fits by contrast. If both are present, the outcome is the original score comparison. All queries remain in the view-level MRR mean, with zero for an absent target.

All selectors receive this same view. Contrast-only maximizes its list-observed pair accuracy; ranking-first maximizes list MRR@10; ranking-constrained contrast maximizes that pair accuracy among fits whose list MRR is at least raw's. The whole-corpus view uses the original pair accuracy. Only scores for documents in the returned list enter this validation comparison. Test pairwise gain always uses the complete original labeled pairs.

Table 16: **What the raw top ten contain.** NevIR columns give answer recall on validation and test, and the fraction of validation pairs for which both queries retrieve both paired documents. NFCorpus columns average over selection halves: recall is the fraction of a query's annotated documents retrieved, and list coverage is the fraction of retrieved documents with a positive qrel. NevIR recall thus counts one answer per query; NFCorpus recall counts the query's supplied positive qrels.

Encoder	NevIR val answer recall	NevIR val paired coverage	NevIR test answer recall	NFCorpus qrel recall	NFCorpus list coverage
BGE-M3	0.896	0.827	0.908	0.142	0.232
BGE-large	0.882	0.849	0.894	0.175	0.278
Multilingual E5	0.938	0.907	0.939	0.138	0.239
mGTE	0.922	0.880	0.919	0.157	0.272
Qwen3-0.6B	0.878	0.840	0.871	0.139	0.223
Qwen3-4B	0.896	0.862	0.900	0.163	0.256

C.3 SELECTION AND THE TWO TEST USES

Table 17 reports all three actual-list sizes and four validation-pair counts. Lists are fixed within a bank. At 225 pairs, all 20 repetitions coincide within a bank, giving 30 bank outcomes. At smaller sizes, the repetitions measure sensitivity to the sampled validation pairs. Sampling seeds depend on list size and view, so different views use their own repeated subsets.

Every chosen readout is evaluated as a global scorer on all test documents and as a reranker of raw's fixed test lists. The fixed top-10 change is recorded for every validation view, even if selection used top-20 or top-50. A separate field evaluates reranking the same k used for selection. Tables 1 and 18, and Figure 3, specify which use is shown. Changes use raw scoring within the same evaluator. Strict score comparisons define NevIR ranks; raw list construction uses a stable sort.

Equation 9 separates raw's target recall from the readout's conditional change where the target is present. Every reranker gives zero reciprocal rank to a missing target, so that recall also bounds list MRR from above. Table 16 reports the observed coverage; Table 18 evaluates the same selected fit globally and within the fixed lists. A positive list change and a negative global change can therefore belong to the same choice.

Table 17: **Actual-list selection with fewer validation pairs.** Cells give the rate of full-test-corpus losses greater than 0.01 MRR@10 / mean test pairwise gain. Twenty validation subsets are averaged within each bank before averaging banks. At 225 pairs these repetitions are identical within a bank. No answer or twin is inserted.

Pairs	List size	Contrast-only	Ranking-first	Ranking-constrained contrast
25	10	84.2% / +0.177	20.5% / +0.096	35.5% / +0.108
25	20	84.8% / +0.177	16.7% / +0.094	31.2% / +0.107
25	50	82.7% / +0.173	11.5% / +0.094	32.2% / +0.111
50	10	87.3% / +0.181	11.2% / +0.104	42.3% / +0.129
50	20	88.5% / +0.180	5.8% / +0.089	26.5% / +0.110
50	50	90.3% / +0.181	4.5% / +0.086	19.3% / +0.108
100	10	94.3% / +0.189	2.3% / +0.102	28.5% / +0.126
100	20	94.3% / +0.190	2.5% / +0.095	25.5% / +0.121
100	50	97.3% / +0.191	1.8% / +0.094	20.3% / +0.119
225	10	100.0% / +0.195	0.0% / +0.099	26.7% / +0.130
225	20	100.0% / +0.194	0.0% / +0.092	16.7% / +0.120
225	50	100.0% / +0.194	0.0% / +0.092	13.3% / +0.120

D READOUT SELECTION WITH GRADED RELEVANCE

D.1 DATA, SPLIT CONSTRUCTION, AND FITTING SIGNAL

The NFCorpus study (recorded as E6) uses the BEIR corpus and qrels at the revisions stored in `nfcorpus_selection.json`. Its grades come from source-link relations: grade 2 for direct links and grade 1 for indirect links (Boteva et al., 2016). The contrast compares these two supplied relevance levels.

The BEIR training qrels contain a single grade. We use the development judgments to obtain two-grade training comparisons. A seeded permutation splits the 324 development queries into equal fitting and selection halves. Each half contains 162 queries; only queries with both grades contribute grade pairs. The fixed 323-query test partition supplies 106 such queries and 16,103 pairs. Across the five runs, the seed changes the development split, initialization, and background draws. Table 19 gives the exact counts.

For each fitting query, an epoch samples at most 64 distinct grade-2/grade-1 pairs from its available combinations. Queries with fewer combinations contribute all of them. This cap limits the influence of queries with many pairs; the number of sampled pairs still sets each query's training weight. The rank-256 lens uses the same 40 epochs, Adam learning rate 10^{-3}, hinge margin $\mu = 0.05$, batch size 128, and 16-member preservation/regularization grid as the polarity bank. The background sampler attempts to exclude annotated documents by resampling collisions. The preservation term compares query-centered background scores, with its mean squared difference divided by the batch's mean squared centered raw score. Corpus and test-query embeddings are read from the existing BEIR cache; development queries are encoded once and cached with their IDs.

Table 18: **Actual top-10 selection by encoder.** Each panel averages five seeds using all validation pairs. Corpus and raw-top-10 columns evaluate the same selected readout in different test settings. Each change is relative to raw in that setting. Adverse counts refer to the full-corpus loss criterion.

Actual raw top-10 validation				
Contrast-only				
Encoder	Adverse	Δpair	Δcorpus MRR	Δtop-10 MRR
BGE-M3	5/5	+0.189	−0.183	−0.085
BGE-large	5/5	+0.147	−0.115	−0.052
Multilingual E5	5/5	+0.207	−0.169	−0.074
mGTE	5/5	+0.194	−0.192	−0.079
Qwen3-0.6B	5/5	+0.194	−0.136	−0.057
Qwen3-4B	5/5	+0.236	−0.311	−0.130
Ranking-first				
Encoder	Adverse	Δpair	Δcorpus MRR	Δtop-10 MRR
BGE-M3	0/5	+0.097	+0.003	+0.007
BGE-large	0/5	+0.089	+0.006	+0.008
Multilingual E5	0/5	+0.019	+0.003	+0.004
mGTE	0/5	+0.127	+0.011	+0.016
Qwen3-0.6B	0/5	+0.107	+0.003	+0.008
Qwen3-4B	0/5	+0.156	+0.008	+0.013
Ranking-constrained contrast				
Encoder	Adverse	Δpair	Δcorpus MRR	Δtop-10 MRR
BGE-M3	5/5	+0.152	−0.035	−0.013
BGE-large	0/5	+0.091	+0.005	+0.008
Multilingual E5	0/5	+0.065	+0.011	+0.012
mGTE	0/5	+0.138	+0.010	+0.016
Qwen3-0.6B	0/5	+0.128	+0.000	+0.007
Qwen3-4B	3/5	+0.209	−0.038	−0.007

Table 19: **NFCorpus query splits for graded preferences.** Each seed splits the 324 development queries into 162 fitting and 162 selection queries. Of the 100 development queries containing both grades, 45–54 enter fitting and 46–55 enter selection. The fixed test set has 323 queries, of which 106 supply 16,103 grade pairs. Pair counts describe all available combinations before the per-epoch sampling cap.

Seed	Fit queries	Fit with both grades	Fit pairs	Selection with both grades	Selection pairs
0	162	45	5,624	55	8,768
1	162	52	6,070	48	8,322
2	162	50	8,739	50	5,653
3	162	49	4,513	51	9,879
4	162	54	9,750	46	4,642

D.2 TWO EVALUATION TARGETS

Let Q_{grade} be the queries with both grades. For each $q \in Q_{\text{grade}}$, let P_q be its ordered pairs (d^+, d^-) with grades 2 and 1. The full-view contrast score is

$$A_{\text{grade}}(h) = \frac{1}{|Q_{\text{grade}}|} \sum_{q \in Q_{\text{grade}}} \frac{1}{|P_q|} \sum_{(d^+,d^-) \in P_q} \mathbf{1}[s_h(q, d^+) > s_h(q, d^-)]. \tag{17}$$

The within-query average prevents a query with many pairs from dominating the reported contrast score. Ranking averages all selection or test queries. nDCG@10 sums $(2^{g(d)} - 1)/\log_2(r + 1)$ over ranks $r = 1, \ldots, 10$ and divides by the ideal DCG from the query's complete qrels. Documents without a qrel receive zero gain in this evaluation.

Actual top-k validation reorders raw's returned lists. In a grade-pair comparison, an absent higher-grade document counts as failure; a present higher-grade document counts as success if the lower-grade one is absent or scores lower. A query's contrast score still averages its full set of grade pairs. Thus both retrieved and missing pairs contribute to the list-observed score. nDCG keeps its ideal denominator from the full qrels. Queries without grade pairs enter ranking means but are excluded from contrast means.

The selector sees 40, 80, or all 162 selection queries. Equations 6–8 apply with the observed grade preference and nDCG@10 in place of NevIR pair accuracy and MRR. Raw remains an eligible candidate and wins ties. A non-raw selection is adverse if its full-test nDCG loss is greater than 0.01. The 0.01 tolerance is measured in each task's own metric: MRR for NevIR and nDCG for NFCorpus. All full-corpus and list changes are paired with their own raw evaluator; stable list sorting and the full-corpus evaluator can differ slightly on tied scores.

Table 20: **NFCorpus selection across query counts and candidate views.** Cells give the full-test-corpus adverse-choice rate / grade-preference gain. All 323 test queries enter nDCG; grade preference averages the 106 queries with both grades. Rates first average 20 subsets within banks. At 162 selection queries the within-bank repetitions coincide.

Queries	Validation list	Contrast-only	Ranking-first	Ranking-constrained contrast
40	Raw top-10	35.5% / −0.002	11.3% / +0.002	9.5% / +0.000
40	Raw top-20	33.2% / −0.001	4.7% / +0.003	6.5% / +0.002
40	Raw top-50	29.7% / −0.001	3.5% / +0.003	2.7% / +0.002
40	Full corpus	27.8% / +0.000	2.8% / +0.003	3.3% / +0.002
80	Raw top-10	36.0% / −0.003	9.3% / +0.002	9.3% / +0.001
80	Raw top-20	26.7% / +0.000	2.3% / +0.003	4.8% / +0.002
80	Raw top-50	23.2% / +0.001	3.5% / +0.003	4.3% / +0.002
80	Full corpus	25.3% / +0.001	3.3% / +0.003	4.0% / +0.003
162	Raw top-10	16.7% / +0.001	6.7% / +0.003	10.0% / +0.001
162	Raw top-20	20.0% / +0.001	6.7% / +0.003	10.0% / +0.001
162	Raw top-50	13.3% / +0.002	6.7% / +0.003	3.3% / +0.002
162	Full corpus	26.7% / +0.001	0.0% / +0.004	6.7% / +0.003

D.3 WHAT THE BANKS AND SELECTED FITS SHOW

The result record contains full-corpus scores, list scores, grade-preference gains, and query-bootstrap intervals for every fit. The largest test grade gain across the 30 banks is 0.0301. Table 21 compares each bank's test-inspected maximum with its validation choice. Figure 4 places all trained fits and the full-corpus ranking-first choices on the same gain axis. Horizontal positions use the recorded values; vertical offsets separate overlapping marks. Nine best-observed fits have intervals wholly above zero. These per-fit intervals are unadjusted for the post-selection search across the bank.

With all selection queries and the full corpus, all three rules choose trained fits in every bank. Ranking-first's lower adverse-choice count is therefore not explained by retaining raw cosine. Mean selected contrast gains are 0.0011 for contrast-only, 0.0037 for ranking-first, and 0.0033 for constrained contrast; corresponding nDCG changes are -0.0089, $+0.0023$, and $+0.0011$. Shorter raw lists

produce different choices (Table 20). Every bank's observed grade gains remain below the 0.05 target.

Table 21: **Available grade-preference gains and validation choices on NFCorpus.** Each bank's best observed value is the largest test grade gain among its 17 readouts, inspected after selection. "Positive interval" counts the unadjusted query-bootstrap intervals above zero for those maxima. These intervals are per-fit. The last column gives the grade gain of the readout chosen by full-corpus ranking-first validation.

Encoder	Raw	Range of best observed gain across five banks	Positive interval	Selected gain
BGE-M3	0.753	$+0.0029$ to $+0.0244$	1/5	$+0.0008$
BGE-large	0.787	$+0.0071$ to $+0.0165$	2/5	$+0.0032$
Multilingual E5	0.758	$+0.0165$ to $+0.0301$	4/5	$+0.0143$
mGTE	0.775	$+0.0052$ to $+0.0218$	2/5	$+0.0065$
Qwen3-0.6B	0.807	$+0.0013$ to $+0.0041$	0/5	$+0.0008$
Qwen3-4B	0.805	$+0.0001$ to $+0.0042$	0/5	-0.0032

NFCorpus and NevIR differ in labels, query averaging, and fitting samples. Raw NFCorpus grade preference is 0.753–0.807. The small gains bound what these fitted candidates achieved, not what all readouts could achieve. They also do not determine whether limited data, the readout family, or the task accounts for the result.

E Candidate expansion and decision paths

A fixed readout assigns the same query–document scores as the pool expands; only the competitors change. This appendix computes expected rank under sampled expansion and separates twin ordering from background competition. The expansion analysis in Section 4.4 uses separately fitted lenses and the external canonical projection. Every query–document score stays fixed as candidates are added.

E.1 From a sampled candidate set to expected MRR

For the rank in Equation 3, let $N = 2{,}079$ be the number of test documents outside a query's answer–twin pair. For readout h, let $t_h(q)$ indicate whether the twin scores above the answer and let $K_h(q)$ count the other documents that do so in the complete corpus. With m background documents drawn uniformly without replacement, the sampled rank is

$$r_h^{(m)}(q) = 1 + t_h(q) + X_h^{(m)}(q), \qquad X_h^{(m)}(q) \sim \mathrm{Hypergeom}(N, K_h(q), m). \tag{18}$$

For a fixed query and readout, the probability mass function and expected reciprocal rank are

$$\begin{aligned} x_- &= \max\{0, m - N + K_h(q)\}, & x_+ &= \min\{m, K_h(q)\}, \\ p_{m,h,q}(x) &= \frac{\binom{K_h(q)}{x}\binom{N-K_h(q)}{m-x}}{\binom{N}{m}}, & x_- \le x &\le x_+, \\ \mathbb{E}[\mathrm{RR@10}_h^{(m)}(q)] &= \sum_{x=x_-}^{x_+} p_{m,h,q}(x)\frac{\mathbf{1}[1 + t_h(q) + x \le 10]}{1 + t_h(q) + x}. \end{aligned} \tag{19}$$

The first factor is the chance of drawing exactly x documents that outrank the answer; the second is reciprocal rank for that draw. Averaging this expectation across queries gives each point in Figure 7. At $m = 0$, the query contributes 1 if the twin does not score higher than the answer and $1/2$ otherwise. At $m = N$, all documents are present and the formula gives the full-corpus score. As in Equation 3, only strictly higher scores count as overtakers; the two-margin accuracy separately requires strict positive margins. The readout stays fixed as m changes. This calculation starts from the full-corpus counts $K_h(q)$ and averages over document draws. The validation study instead uses the visible scores to choose among readouts.

Reading the rank decomposition. Correcting the twin's order can remove at most one document from above the answer. Other documents may still outrank it, so a better paired decision need not give a better corpus rank. Equations 18–19 count this competition under uniform sampling. They average over possible background draws with the readout and all scores fixed; they do not describe retraining as the document pool grows.

Relation to validation selection. The expansion calculation starts from known full-corpus overtaker counts. A selector with a restricted validation view does not observe all of those counts. Appendix B.8 considers estimating counts from sampled documents; applying reciprocal rank afterward is nonlinear. The expansion curves thus explain how competitors affect evaluation, while the selection experiment separately measures whether the available validation scores identify a useful readout.

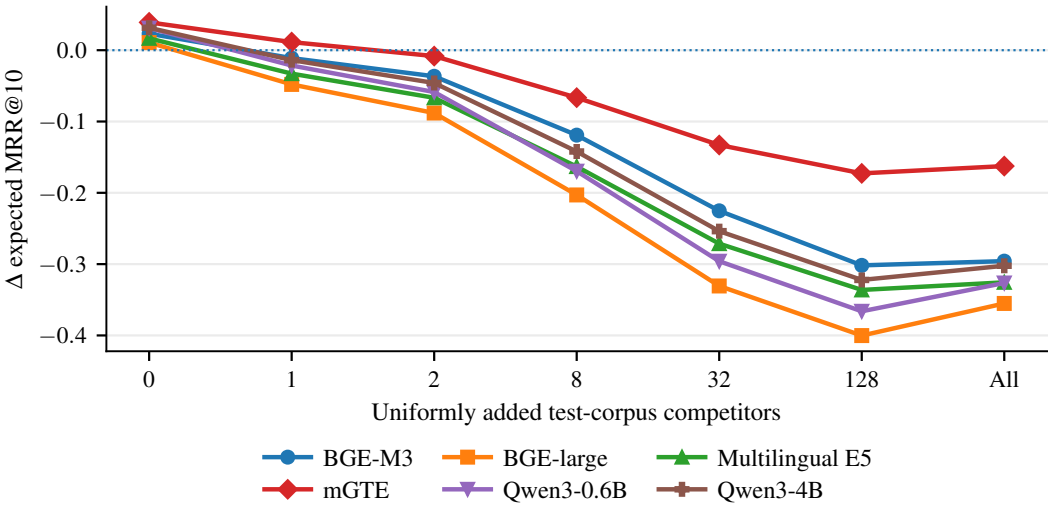


Figure 7: **Expected ranking change under fixed-score candidate expansion.** The answer and twin remain present while additional documents are sampled uniformly from the other 2,079 test-corpus documents. Each curve is the two-sided canonical projection's change in expected MRR@10 from raw cosine for one encoder. Queries, labels, vectors, and all query–document scores remain fixed. The exact calculation averages over the possible background draws; "All" uses the complete background pool.

E.2 Two margins, three possible outcomes

For a fixed pair and readout h, abbreviate $m_i = m_i(h)$ from Equation 2, and define

$$a = \frac{m_1 + m_2}{2}, \qquad b = \frac{m_1 - m_2}{2}, \qquad \gamma = a - |b| = \min(m_1, m_2). \tag{20}$$

The two margins are $a + b$ and $a - b$. Excluding ties, both queries are correct when $a > |b|$, neither is correct when $a < -|b|$, and exactly one is correct when $|a| < |b|$. For a cosine score, writing q_i, d_i for the vectors actually scored, $a = \frac{1}{2}(q_1 - q_2)^\top (d_1 - d_2)$ and $b = \frac{1}{2}(q_1 + q_2)^\top (d_1 - d_2)$. The first term measures the contrast associated with the query change; the second measures the document preference shared by the two queries. For a lens, q_i and d_i denote the transformed, normalized vectors. Table 22 reports these quantities for the readouts and the external-span removal. The transition records identify the pairs that cross each decision boundary. Under raw cosine 0.148–0.267 of the pairs have both queries correct, 0.693–0.829 exactly one, and 0.023–0.040 neither; the both-sides projection moves these shares to 0.305–0.408, 0.505–0.612, and 0.076–0.106. Both completely correct and completely wrong pairs become more frequent (pair-bootstrap 95% intervals for the latter increase exclude zero on six of six encoders).

The mean query-conditioned component a rises from $+0.001$–$+0.004$ to $+0.014$–$+0.030$ and the mean shared document preference $|b|$ from 0.004–0.010 to 0.023–0.036; on the query side alone a reaches $+0.002$–$+0.008$ and $|b|$ stays at 0.005–0.011, with 0.242–0.386 of pairs both correct.

Table 22: Pair outcomes on the 1,383 NevIR test pairs. For a pair with margins $m_1 = s(q_1, d_1) - s(q_1, d_2)$ and $m_2 = s(q_2, d_2) - s(q_2, d_1)$, let $a = (m_1 + m_2)/2$ (the query-conditioned flip) and $b = (m_1 - m_2)/2$ (the document preference shared by both queries); excluding ties, both queries are correct iff $a > |b|$, exactly one iff $|a| < |b|$, neither iff $a < -|b|$. Cells give the shares both / one / neither under raw cosine, the canonical projection on both sides and on the query only, and the W_{doc} lens; the last two columns give the mean of a and of $|b|$, raw $\rightarrow$ projected.

Shares both / one / neither correct: raw and two-sided projection

Model	Raw cosine	Canonical, both sides
BGE-M3	0.267 / 0.693 / 0.040	0.408 / 0.505 / 0.087
BGE-large	0.256 / 0.711 / 0.033	0.371 / 0.523 / 0.106
multilingual E5	0.188 / 0.780 / 0.032	0.305 / 0.612 / 0.082
mGTE	0.148 / 0.829 / 0.023	0.372 / 0.537 / 0.091
Qwen3-0.6B	0.153 / 0.816 / 0.031	0.338 / 0.570 / 0.092
Qwen3-4B	0.200 / 0.776 / 0.025	0.378 / 0.546 / 0.076

Shares both / one / neither correct: query-only projection and lens

Model	Canonical, query only	W_{doc}
BGE-M3	0.386 / 0.549 / 0.065	0.449 / 0.458 / 0.093
BGE-large	0.351 / 0.582 / 0.067	0.401 / 0.519 / 0.080
multilingual E5	0.242 / 0.712 / 0.046	0.392 / 0.518 / 0.090
mGTE	0.269 / 0.686 / 0.045	0.345 / 0.591 / 0.064
Qwen3-0.6B	0.291 / 0.650 / 0.059	0.385 / 0.515 / 0.100
Qwen3-4B	0.304 / 0.656 / 0.040	0.420 / 0.488 / 0.092

Mean score components: raw to two-sided projection

Model	Mean a	Mean $\lvert b\rvert$
BGE-M3	$+0.004 \rightarrow +0.025$	$0.009 \rightarrow 0.030$
BGE-large	$+0.003 \rightarrow +0.030$	$0.008 \rightarrow 0.036$
multilingual E5	$+0.001 \rightarrow +0.014$	$0.004 \rightarrow 0.023$
mGTE	$+0.002 \rightarrow +0.026$	$0.009 \rightarrow 0.031$
Qwen3-0.6B	$+0.002 \rightarrow +0.019$	$0.008 \rightarrow 0.034$
Qwen3-4B	$+0.003 \rightarrow +0.025$	$0.010 \rightarrow 0.029$

Individual transitions explain how these averages arise. Among pairs that fall from one correct to neither, a was already negative before the projection on 0.600–0.729 of them and falls further by -0.046–-0.027 while $|b|$ changes by $+0.002$–$+0.005$, so the negative query-conditioned margin becomes more negative. Among pairs that fall from both correct to one, $|b|$ grows by $+0.032$–$+0.051$ against $+0.008$–$+0.016$ for a, so the shared document preference outweighs the correct contrast. Pairs gained from one correct to both raise a by $+0.028$–$+0.057$. These transitions include both larger correct margins and larger errors. Changes in the two components must be read against each pair's decision boundary.

Removing P_{ext} from the output of the validation-selected lens changes the mean of a by -0.009–-0.005 and of $|b|$ by -0.005–-0.004 (the fixed fit of Appendix H: -0.009–-0.004 and -0.005–-0.004), and the both-correct share by -0.046–$+0.004$: the removal lowers the average query-conditioned contrast but also changes the shared preference, and the accuracy effect depends on both relative to each pair's boundary.

Grouping the document pairs by their token differences (explicit-cue change 534, single-word swap 171, other edits 678 pairs) shows the projection's gain in every class: both-correct shares move from 0.165–0.311 to 0.361–0.455 on cue changes, from 0.123–0.240 to 0.234–0.363 on single-word swaps, and from 0.112–0.252 to 0.277–0.398 on other edits. Equation 20 distinguishes the transitions: decreasing a and increasing $|b|$ can cross different decision boundaries. The application-side comparison below examines how query-only scoring changes this balance.

E.3 THE TWIN AND THE REST OF THE CORPUS

A relevant document competes with its twin and with every other candidate. Table 23 follows both parts of its rank on the same queries. Applied to every stored vector, the both-sides projection lowers the twin penalty from 0.096–0.113 to 0.029–0.055 on every encoder while the mean number of non-twin documents above the relevant one rises from 5.2–14.0 to 167.8–337.7. It corrects 389–539 twin orders and creates 261–405 new twin errors. The lower twin penalty and greater background overtaking account for the net MRR@10 loss (Table 23). The tie-break keeps the non-twin count at 5.1–14.0 and corrects 325–453 twin orders against 207–367 new errors. The local policy in Appendix F limits the allowed exchanges; the remaining documents still compete within those limits.

Table 23: From the pair decision to corpus rank on the NevIR test corpus ($2{,}766$ queries against $2{,}081$ documents). Twin penalty is the MRR@10 lost to the negated twin (MRR@10 with the twin removed minus MRR@10); other above is the mean number of non-twin documents ranked above the relevant document. Cells give raw cosine / both-sides projection / near-duplicate tie-break. Corrections and new errors count queries whose relevant document moves above, or below, its twin relative to raw cosine, for the both-sides projection applied to every vector, the full top-10 reorder, and the tie-break.

(a) Scores and competitors: raw / projection / tie-break

Model	MRR@10	Twin penalty	Other above
BGE-M3	0.577 / 0.282 / 0.598	0.100 / 0.036 / 0.085	8.23 / 201.14 / 8.21
BGE-large	0.561 / 0.206 / 0.572	0.096 / 0.030 / 0.094	10.43 / 337.70 / 10.39
multilingual E5	0.591 / 0.265 / 0.610	0.113 / 0.040 / 0.105	5.16 / 298.59 / 5.14
mGTE	0.556 / 0.393 / 0.584	0.109 / 0.055 / 0.091	7.52 / 167.76 / 7.48
Qwen3-0.6B	0.528 / 0.201 / 0.546	0.105 / 0.029 / 0.092	13.99 / 315.76 / 13.97
Qwen3-4B	0.567 / 0.263 / 0.589	0.102 / 0.038 / 0.093	10.85 / 261.11 / 10.82

(b) Corrected / newly incorrect twin orders

Model	Full corpus	Top-10	Tie-break
BGE-M3	389 / 261	352 / 233	325 / 207
BGE-large	446 / 387	386 / 351	361 / 333
multilingual E5	497 / 405	459 / 373	453 / 367
mGTE	528 / 312	472 / 286	444 / 269
Qwen3-0.6B	539 / 368	465 / 317	445 / 305
Qwen3-4B	487 / 311	431 / 278	405 / 256

E.4 QUERY-ONLY AND TWO-SIDED APPLICATION

For $T(z) = \mathrm{norm}(z + \alpha P^\top P z)$, transforming both sides weights the cross term $z_q^\top P^\top P z_d$ by $2\alpha + \alpha^2$ before renormalization. Query-only application weights it by α and leaves document vectors unchanged. Table 24 compares these methods at matched cross-term weight and at matched pairwise gain. At the matched query scale $2\alpha + \alpha^2$, query-only accuracy is 0.244–0.442 versus 0.305–0.408 for both sides at α; corpus MRR is 0.050–0.183 versus 0.201–0.393. For BGE-M3 and BGE-large, query-only scoring reaches the two-sided pairwise gain at $\alpha_q = 6.0$–23.0. Its MRR there is 0.273–0.469 versus 0.206–0.282. NegConstraint and SciFact changes are -0.078 to -0.039 and -0.350 to -0.138, respectively, versus -0.244 to -0.181 and -0.464 to -0.305 for both sides. On the other four encoders, query-only scoring reaches only 61–95% of the two-sided gain through $\alpha = 32$. Thus, at matched cross-term weight, query-only scoring loses more corpus MRR. Where pairwise gains can be matched, it retains more MRR and incurs smaller transfer costs. The two comparisons hold different quantities fixed: the score coefficient before normalization, or the observed accuracy improvement. Two-sided application also changes each document's normalization.

Table 24: Application side at matched amplification. With $T(z) = \mathrm{norm}(z + \alpha P^\top P z)$ on both sides, the cross term of the score is weighted by $2\alpha + \alpha^2$, whereas query-only application weights it by α; both-sides application also renormalizes every document. Panel (a): the both-sides projection at the selected scale (pairwise accuracy, corpus MRR@10, nDCG@10 change on NegConstraint / SciFact). Panel (b): query-only application at the cross-term-matched scale $2\alpha + \alpha^2$. Panel (c): query-only application at the scale that matches the both-sides pairwise gain (interpolated on a 14-point grid), with its corpus MRR@10 and costs on the other tasks.

(a) Two-sided projection at the selected scale

Model	α	Pairwise	MRR	NegC. / SciFact
BGE-M3	4	0.408	0.282	−0.244 / −0.305
BGE-large	8	0.371	0.206	−0.181 / −0.464
multilingual E5	8	0.305	0.265	−0.313 / −0.493
mGTE	8	0.372	0.393	−0.062 / −0.368
Qwen3-0.6B	8	0.338	0.201	−0.352 / −0.376
Qwen3-4B	8	0.378	0.263	−0.329 / −0.285

(b) Query-only projection, matched cross-term weight

Model	α	Pairwise	MRR
BGE-M3	24	0.442	0.183
BGE-large	80	0.395	0.050
multilingual E5	80	0.244	0.056
mGTE	80	0.357	0.168
Qwen3-0.6B	80	0.339	0.076
Qwen3-4B	80	0.377	0.118

(c) Query-only projection, matched pairwise gain

Model	α	Pairwise	MRR	NegC. / SciFact
BGE-M3	6.0	0.408	0.469	−0.039 / −0.138
BGE-large	23.0	0.372	0.273	−0.078 / −0.350
multilingual E5	–	–	–	–
mGTE	–	–	–	–
Qwen3-0.6B	–	–	–	–
Qwen3-4B	–	–	–	–

E.5 Expansion curves across readouts and datasets

Tables 25 and 26 give the complete expansion curves introduced in this appendix for every readout on the NevIR test corpus and on ExcluIR, with pair-clustered intervals at $m = 0$ and $m = 1$. The two additional tasks give different curves. CondaQA uses the original passage as query, its paraphrase as the relevant document, and its negation-flipped edit as the twin. Under the two-sided projection, almost no background passage outranks the paraphrase. The MRR change stays within 0.002 of its two-candidate value ($+0.041$ to $+0.064$ across encoders) as the candidate set expands to the whole corpus. ExcluIR tests exclusion instead. Its polarity readouts already score below raw at $m = 0$ on all six encoders (-0.042 to -0.014), and remain below raw on the whole corpus (-0.462 to -0.074; Table 26).

Table 25: Candidate expansion on the NevIR test corpus, every readout. Change in expected MRR@10 against raw cosine at each number m of uniformly drawn extra candidates (brackets: pair-clustered bootstrap 95% intervals at $m = 0$ and $m = 1$), and the smallest m with a negative change (never: none up to the whole corpus). Every readout is fitted or selected as elsewhere and applied unchanged.

Readout	$m = 0$	1	2	8	32	128	All	m^*
BGE-M3								
canonical both sides	$+0.023$ $[+0.015, +0.031]$	-0.011 $[-0.019, -0.003]$	-0.037	-0.119	-0.225	-0.302	-0.296	1
canonical query only	$+0.023$ $[+0.017, +0.030]$	$+0.021$ $[+0.014, +0.027]$	$+0.019$	$+0.009$	-0.010	-0.036	-0.063	32
W_{doc}	$+0.034$ $[+0.025, +0.042]$	$+0.013$ $[+0.004, +0.021]$	-0.003	-0.055	-0.122	-0.173	-0.181	2
external mean difference	$+0.018$ $[+0.009, +0.026]$	-0.053 $[-0.062, -0.044]$	-0.101	-0.230	-0.357	-0.409	-0.356	1
Semantic Adapter	$+0.003$ $[-0.004, +0.009]$	-0.003 $[-0.010, +0.003]$	-0.008	-0.024	-0.048	-0.074	-0.093	1
BGE-large								
canonical both sides	$+0.011$ $[+0.002, +0.019]$	-0.048 $[-0.057, -0.039]$	-0.088	-0.203	-0.331	-0.400	-0.355	1
canonical query only	$+0.015$ $[+0.008, +0.022]$	$+0.012$ $[+0.005, +0.018]$	$+0.009$	-0.002	-0.020	-0.044	-0.068	8
W_{doc}	$+0.025$ $[+0.018, +0.033]$	$+0.013$ $[+0.005, +0.020]$	$+0.003$	-0.028	-0.071	-0.107	-0.122	4
external mean difference	$+0.014$ $[+0.006, +0.023]$	-0.036 $[-0.046, -0.027]$	-0.071	-0.171	-0.277	-0.329	-0.282	1
Semantic Adapter	-0.003 $[-0.009, +0.003]$	-0.006 $[-0.012, -0.000]$	-0.008	-0.017	-0.031	-0.046	-0.061	0
multilingual E5								
canonical both sides	$+0.017$ $[+0.008, +0.025]$	-0.033 $[-0.041, -0.024]$	-0.067	-0.163	-0.271	-0.336	-0.326	1
canonical query only	$+0.010$ $[+0.004, +0.016]$	$+0.007$ $[+0.001, +0.014]$	$+0.005$	-0.004	-0.022	-0.046	-0.078	8
W_{doc}	$+0.039$ $[+0.030, +0.047]$	$+0.020$ $[+0.012, +0.028]$	$+0.006$	-0.042	-0.107	-0.157	-0.169	4
external mean difference	$+0.012$ $[+0.004, +0.021]$	-0.030 $[-0.039, -0.021]$	-0.058	-0.134	-0.213	-0.257	-0.237	1
Semantic Adapter	-0.014 $[-0.021, -0.007]$	-0.043 $[-0.051, -0.035]$	-0.064	-0.127	-0.212	-0.284	-0.312	0
mGTE								
canonical both sides	$+0.039$ $[+0.031, +0.048]$	$+0.011$ $[+0.003, +0.020]$	-0.008	-0.067	-0.133	-0.173	-0.163	2
canonical query only	$+0.025$ $[+0.019, +0.031]$	$+0.024$ $[+0.018, +0.030]$	$+0.024$	$+0.021$	$+0.015$	$+0.007$	-0.002	512
W_{doc}	$+0.040$ $[+0.033, +0.048]$	$+0.035$ $[+0.028, +0.042]$	$+0.031$	$+0.016$	-0.008	-0.033	-0.050	32
external mean difference	$+0.038$ $[+0.030, +0.047]$	$+0.010$ $[+0.002, +0.019]$	-0.009	-0.068	-0.136	-0.177	-0.168	2
Semantic Adapter	$+0.002$ $[-0.003, +0.006]$	$+0.001$ $[-0.003, +0.005]$	$+0.000$	-0.003	-0.010	-0.019	-0.027	4
Qwen3-0.6B								
canonical both sides	$+0.031$ $[+0.023, +0.039]$	-0.021 $[-0.030, -0.014]$	-0.059	-0.170	-0.296	-0.366	-0.326	1
canonical query only	$+0.028$ $[+0.021, +0.034]$	$+0.025$ $[+0.018, +0.032]$	$+0.023$	$+0.014$	-0.002	-0.022	-0.048	32
W_{doc}	$+0.038$ $[+0.029, +0.048]$	-0.065 $[-0.074, -0.056]$	-0.132	-0.310	-0.478	-0.534	-0.436	1
external mean difference	$+0.028$ $[+0.020, +0.037]$	-0.019 $[-0.028, -0.011]$	-0.052	-0.149	-0.254	-0.305	-0.264	1
Semantic Adapter	-0.003 $[-0.007, +0.003]$	-0.007 $[-0.012, -0.002]$	-0.010	-0.021	-0.034	-0.045	-0.049	0

Continued on the next page

Table 25 (continued)

Readout	$m = 0$	1	2	8	32	128	All	m^*
Qwen3-4B								
canonical both sides	$+0.031$ $[+0.023, +0.040]$	-0.014 $[-0.022, -0.006]$	-0.046	-0.142	-0.254	-0.322	-0.302	1
canonical query only	$+0.023$ $[+0.017, +0.029]$	$+0.021$ $[+0.015, +0.027]$	$+0.020$	$+0.014$	$+0.004$	-0.009	-0.022	64
W_{doc}	$+0.038$ $[+0.029, +0.047]$	-0.063 $[-0.072, -0.054]$	-0.130	-0.305	-0.473	-0.537	-0.456	1
external mean difference	$+0.025$ $[+0.018, +0.033]$	-0.017 $[-0.025, -0.009]$	-0.046	-0.133	-0.231	-0.287	-0.263	1
Semantic Adapter	$+0.003$ $[-0.001, +0.008]$	$+0.000$ $[-0.004, +0.005]$	-0.002	-0.011	-0.023	-0.032	-0.028	2

Table 26: Candidate expansion on ExcluIR (query, satisfying and violating document; the other documents as candidates; every readout applied unchanged, so W_{doc} is a transfer). Change in expected MRR@10 against raw cosine at each number m of uniformly drawn extra candidates (brackets: pair-clustered bootstrap 95% intervals at $m = 0$ and $m = 1$), and the smallest m with a negative change (never: none up to the whole corpus). Every readout is fitted or selected as elsewhere and applied unchanged.

Readout	$m = 0$	1	2	8	32	128	All	m^*
BGE-M3								
canonical both sides	-0.016 $[-0.021, -0.011]$	-0.016 $[-0.022, -0.011]$	-0.017	-0.020	-0.026	-0.039	-0.097	0
canonical query only	-0.002 $[-0.004, +0.000]$	-0.002 $[-0.004, +0.000]$	-0.002	-0.002	-0.002	-0.002	-0.004	0
W_{doc}	-0.009 $[-0.014, -0.004]$	-0.009 $[-0.014, -0.004]$	-0.010	-0.011	-0.014	-0.019	-0.035	0
external mean difference	-0.017 $[-0.024, -0.011]$	-0.022 $[-0.029, -0.015]$	-0.026	-0.042	-0.072	-0.107	-0.175	0
Semantic Adapter	-0.017 $[-0.023, -0.012]$	-0.018 $[-0.023, -0.013]$	-0.018	-0.020	-0.025	-0.031	-0.050	0
BGE-large								
canonical both sides	-0.021 $[-0.029, -0.014]$	-0.025 $[-0.033, -0.018]$	-0.029	-0.042	-0.066	-0.094	-0.172	0
canonical query only	$+0.000$ $[-0.003, +0.004]$	$+0.000$ $[-0.003, +0.004]$	$+0.000$	-0.000	-0.000	-0.001	-0.005	8
W_{doc}	-0.002 $[-0.007, +0.003]$	-0.002 $[-0.007, +0.003]$	-0.002	-0.003	-0.005	-0.009	-0.024	0
external mean difference	-0.018 $[-0.025, -0.011]$	-0.021 $[-0.029, -0.014]$	-0.024	-0.036	-0.058	-0.083	-0.147	0
Semantic Adapter	-0.008 $[-0.013, -0.003]$	-0.008 $[-0.013, -0.003]$	-0.008	-0.009	-0.010	-0.011	-0.023	0
multilingual E5								
canonical both sides	-0.020 $[-0.028, -0.012]$	-0.030 $[-0.038, -0.022]$	-0.038	-0.065	-0.107	-0.148	-0.236	0
canonical query only	-0.001 $[-0.005, +0.003]$	-0.001 $[-0.005, +0.003]$	-0.001	-0.002	-0.002	-0.003	-0.011	0
W_{doc}	-0.001 $[-0.007, +0.004]$	-0.002 $[-0.007, +0.004]$	-0.002	-0.005	-0.010	-0.018	-0.047	0
external mean difference	-0.014 $[-0.021, -0.006]$	-0.020 $[-0.027, -0.013]$	-0.025	-0.042	-0.068	-0.094	-0.151	0
Semantic Adapter	-0.022 $[-0.031, -0.014]$	-0.032 $[-0.041, -0.024]$	-0.040	-0.065	-0.106	-0.148	-0.260	0

Continued on the next page

Table 26 (continued)

Readout	$m = 0$	1	2	8	32	128	All	m^*
mGTE								
canonical both sides	-0.014 $[-0.018, -0.009]$	-0.015 $[-0.020, -0.010]$	-0.017	-0.022	-0.032	-0.044	-0.074	0
canonical query only	$+0.000$ $[-0.002, +0.002]$	$+0.000$ $[-0.002, +0.002]$	$+0.000$	$+0.000$	$+0.000$	-0.000	-0.000	128
W_{doc}	-0.002 $[-0.005, +0.001]$	-0.002 $[-0.005, +0.001]$	-0.002	-0.003	-0.003	-0.004	-0.007	0
external mean difference	-0.014 $[-0.019, -0.009]$	-0.015 $[-0.020, -0.011]$	-0.017	-0.023	-0.033	-0.046	-0.076	0
Semantic Adapter	$+0.000$ $[-0.002, +0.003]$	$+0.000$ $[-0.002, +0.003]$	$+0.000$	-0.000	-0.001	-0.001	-0.002	8
Qwen3-0.6B								
canonical both sides	-0.042 $[-0.050, -0.035]$	-0.053 $[-0.061, -0.046]$	-0.063	-0.104	-0.184	-0.284	-0.462	0
canonical query only	-0.002 $[-0.005, +0.001]$	-0.002 $[-0.005, +0.001]$	-0.002	-0.002	-0.002	-0.003	-0.006	0
W_{doc}	-0.077 $[-0.087, -0.069]$	-0.126 $[-0.136, -0.117]$	-0.162	-0.276	-0.419	-0.512	-0.608	0
external mean difference	-0.040 $[-0.047, -0.034]$	-0.049 $[-0.056, -0.043]$	-0.057	-0.092	-0.163	-0.252	-0.401	0
Semantic Adapter	-0.002 $[-0.005, +0.002]$	-0.002 $[-0.006, +0.001]$	-0.002	-0.003	-0.004	-0.006	-0.011	0
Qwen3-4B								
canonical both sides	-0.030 $[-0.037, -0.024]$	-0.037 $[-0.044, -0.030]$	-0.043	-0.069	-0.126	-0.212	-0.404	0
canonical query only	-0.001 $[-0.003, +0.001]$	-0.001 $[-0.003, +0.001]$	-0.001	-0.001	-0.001	-0.002	-0.004	0
W_{doc}	-0.056 $[-0.065, -0.048]$	-0.097 $[-0.106, -0.088]$	-0.128	-0.234	-0.371	-0.468	-0.578	0
external mean difference	-0.033 $[-0.039, -0.027]$	-0.039 $[-0.046, -0.033]$	-0.045	-0.072	-0.132	-0.220	-0.392	0
Semantic Adapter	$+0.003$ $[-0.000, +0.006]$	$+0.003$ $[-0.000, +0.006]$	$+0.003$	$+0.002$	$+0.001$	$+0.000$	-0.006	256

E.6 Scores over the full document collection

Pairwise accuracy scores each query against its own two documents. To test the same score beyond the pair, we rank each NevIR query's designated answer among all unique test documents, which adds documents outside the designated contrast pair. We report MRR@10, recall at 1 (R@1), twin precedence (the relevant document outranks its negated twin), and the non-twin MRR obtained after removing the twin from the corpus, which separates the effect of non-twin competition. Readouts are raw cosine, the canonical both-sides projection, W_{doc}, and a query-only projection at $\alpha = 1$; the full scale-by-side sweep is reported in the supplementary results. Because the closed-form projection uses no NevIR examples for construction or selection, it is also scored on all $2{,}556$ NevIR pairs.

Reading the corpus columns. The twin-precedence column records the designated contrast, while MRR@10 also counts the other documents ahead of the answer. Removing the twin leaves a separate measure of those background ranks.

E.7 Alternative scores on the same vectors

The lens and the canonical projection are low-rank metric readouts, so classical alternatives are scored on the same cached vectors: centered and shrinkage-whitened cosine, a closed-form rank-1 mean-difference direction and a shrinkage discriminant direction (both amplified on both sides with the scale selected on Jina triplet AUC, as for the canonical axis), a symmetric low-rank metric $s(q, d) = T(q)^\top T(d)$ with $T(z) = \text{norm}(z + WW^\top z)$ fitted on NevIR train with the lens loss, and W_{doc}.

Table 27: Full-corpus NevIR evaluation: each test query ranks its benchmark-designated answer among all 2,081 test documents (2,766 queries). Cells give MRR@10 / twin precedence, where twin precedence is the probability that the relevant document outranks its negated twin. Closed form is the canonical both-sides projection at the externally selected scale; query-only uses $\alpha = 1$ on the query side. Non-twin MRR removes the twin from the corpus and measures its rank change after two-sided projection. The last column is the closed form on all 2,556 NevIR pairs (pairwise accuracy).

(a) Corpus MRR@10 / twin precedence

Model	Raw	Closed form (both)	W_{doc}
BGE-M3	0.577 / 0.614	0.282 / 0.660	0.394 / 0.678
BGE-large	0.561 / 0.611	0.206 / 0.633	0.446 / 0.660
multilingual E5	0.591 / 0.578	0.265 / 0.611	0.422 / 0.651
mGTE	0.556 / 0.562	0.393 / 0.640	0.501 / 0.641
Qwen3-0.6B	0.528 / 0.561	0.201 / 0.623	0.098 / 0.643
Qwen3-4B	0.567 / 0.587	0.263 / 0.651	0.121 / 0.664

(b) Query-side application and complementary evaluations

Model	Query-only $\alpha = 1$	Non-twin MRR raw to projected	All-pairs accuracy raw to projected
BGE-M3	0.575 / 0.637	0.678 → 0.317	0.268 → 0.404
BGE-large	0.567 / 0.625	0.657 → 0.236	0.256 → 0.377
multilingual E5	0.596 / 0.584	0.703 → 0.305	0.193 → 0.315
mGTE	0.562 / 0.580	0.664 → 0.448	0.156 → 0.354
Qwen3-0.6B	0.528 / 0.573	0.633 → 0.230	0.161 → 0.328
Qwen3-4B	0.569 / 0.596	0.669 → 0.301	0.196 → 0.365

Table 28: Readouts under one protocol on all six encoders (ranges over encoders). Closed-form rows use the synthetic inventory for direction and Jina triplet AUC for operating-point selection; centered and whitened cosine use the NevIR train mean and covariance; the symmetric low-rank metric and W_{doc} are fitted on NevIR train with the same hinge loss. NevIR and CondaQA are pairwise accuracies; NegConstraint and SciFact are nDCG@10 with the readout on both sides; corpus is the NevIR corpus contract (MRR@10 / twin precedence).

(a) Pairwise accuracy

Readout	NevIR	CondaQA
raw cosine	0.148–0.267	0.594–0.685
centered cosine	0.146–0.299	0.593–0.688
whitened cosine	0.106–0.171	0.573–0.635
mean-difference, rank 1	0.307–0.399	0.712–0.766
discriminant, rank 1	0.188–0.371	0.660–0.722
canonical projection	0.305–0.408	0.722–0.775
symmetric low-rank metric	0.318–0.428	0.701–0.792
residual lens W_{doc}	0.345–0.449	0.714–0.788

(b) Other contracts and corpus rank

Readout	NegConstraint	SciFact	Corpus MRR	Twin
raw cosine	0.674–0.813	0.641–0.739	0.528–0.591	0.561–0.614
centered cosine	0.637–0.808	0.635–0.755	0.540–0.594	0.560–0.629
whitened cosine	0.599–0.782	0.435–0.739	0.484–0.548	0.541–0.572
mean-difference, rank 1	0.393–0.747	0.254–0.617	0.222–0.387	0.603–0.650
discriminant, rank 1	0.527–0.816	0.498–0.738	0.370–0.565	0.570–0.640
canonical projection	0.360–0.750	0.196–0.454	0.201–0.393	0.611–0.660
symmetric low-rank metric	0.060–0.266	0.048–0.217	0.039–0.114	0.619–0.667
residual lens W_{doc}	0.244–0.787	0.191–0.631	0.098–0.501	0.641–0.678

Several polarity-directed scores improve contrast accuracy while lowering corpus MRR. The mean-difference direction gives results close to the canonical projection on NevIR. These alternative readouts were fitted or selected separately from the main banks.

F LOCAL RERANKING AND CONDITIONAL EFFECTS

A local policy must both form candidate groups and choose an order inside them. We vary the grouping rule and the within-group scorer separately, then identify the queries whose evaluation score can change. Each group retains its original rank slots; a member may still move past a document from another group.

Two forms of reranking. Section 4.4 reorders every document in raw's fixed top-k list. The local policies below are more restrictive: they form groups and reorder documents only among each group's existing rank slots. The list membership remains fixed in both cases. Consequently, neither procedure can retrieve an answer missing from that list, but the local policy also limits which retrieved documents may exchange positions.

What these controls test. Holding groups fixed while changing the within-group scorer tests the ordering rule. Changing group membership while retaining the scorer tests the grouping rule. Tables 29–31 report these comparisons. They examine local reranking behavior, not the choice among the preservation-trained banks used in Tables 1–3. Each reported change should therefore be read with its own candidate list, fitting protocol, and raw baseline.

Table 29: Where a polarity readout may be consulted. The canonical both-sides projection is applied to every stored vector (full), only to reorder each query's raw top-10 (top-10), or only to order near-duplicate candidates inside the raw top-10 (tie-break: candidates whose pairwise cosine is at least 0.98 are ordered by the readout, every other candidate keeps its raw position). NevIR cells give corpus-level MRR@10 / twin precedence; the tie-break intervals are 95% bootstrap intervals that resample document pairs with both of their queries; the last two columns give the nDCG@10 change on NegConstraint and SciFact for top-10 / tie-break.

Model	Raw	Full	Top-10	Tie-break
BGE-M3	0.577 / 0.614	0.282 / 0.660	0.427 / 0.657	0.598 / 0.657
BGE-large	0.561 / 0.611	0.206 / 0.633	0.384 / 0.624	0.572 / 0.621
multilingual E5	0.591 / 0.578	0.265 / 0.611	0.421 / 0.609	0.610 / 0.609
mGTE	0.556 / 0.562	0.393 / 0.640	0.474 / 0.629	0.584 / 0.625
Qwen3-0.6B	0.528 / 0.561	0.201 / 0.623	0.375 / 0.615	0.546 / 0.612
Qwen3-4B	0.567 / 0.587	0.263 / 0.651	0.424 / 0.643	0.589 / 0.641

Model	Tie ΔMRR [95% CI]	Tie Δtwin [95% CI]	NegConstraint Δ	SciFact Δ
BGE-M3	+0.020 [+0.011, +0.029]	+0.043 [+0.028, +0.057]	−0.103 / +0.001	−0.147 / +0.000
BGE-large	+0.010 [+0.001, +0.020]	+0.010 [−0.006, +0.027]	−0.077 / −0.000	−0.216 / +0.000
multilingual E5	+0.019 [+0.008, +0.029]	+0.031 [+0.015, +0.046]	−0.139 / −0.000	−0.244 / +0.000
mGTE	+0.029 [+0.019, +0.039]	+0.063 [+0.047, +0.079]	−0.037 / +0.000	−0.180 / +0.000
Qwen3-0.6B	+0.019 [+0.009, +0.028]	+0.051 [+0.036, +0.066]	−0.170 / −0.000	−0.169 / +0.000
Qwen3-4B	+0.023 [+0.013, +0.032]	+0.054 [+0.039, +0.069]	−0.175 / +0.000	−0.145 / +0.000

F.1 CANDIDATE GROUPS AND CONTROL COMPARISONS

For each query, we connect raw top-ten documents with pairwise cosine of at least 0.98, take the connected components, and reorder each component within its original slots. Table 29 compares this policy with global scoring and full top-ten reranking.

Table 30 reports group sizes, permitted within-group pairs, and final-ranking inversions. Both matched controls score below raw cosine: one randomizes order within the original groups; the other assigns candidates to random groups of the same sizes before applying the original scorer. Group sizes are matched, while the resulting inversion counts are measured separately.

Other scorers can also work within the cosine groups. The lens stays within 0.01 of the canonical direction, and the external mean-difference direction beats raw on every encoder and the projection on two. The leading NevIR-train difference direction stays at or below raw, while a random direction leaves MRR@10 near raw.

Table 30: What the near-duplicate gate does (raw top-10; connected components of the cosine ≥ 0.98 link, each reordered inside its own slots by the canonical scorer, with singletons fixed). Trigger: share of queries with a component of size ≥ 2; change: share whose order actually changes; Kendall inversions and within-group pairs are reported as two separate counts; chained: share of groups containing a member pair below the threshold (single-link chaining); interleaved: share of groups whose members occupy nonconsecutive raw slots; in top / same group: share of queries whose relevant document and twin are both in the raw top-10, and in the same group. MRR@10 of the tie-break is compared with two matched controls (the same groups in random order; random candidate subsets of the same sizes reordered by the same scorer; means over 30 draws) and with two oracles (relevant document first inside its group; relevant document first among the top-10).

Group activity and size

Model	Trigger	Change	Kendall inversions / within-group pairs	Mean group size
BGE-M3	1.000	0.856	8,073 / 20,877	2.73
BGE-large	1.000	0.922	10,196 / 21,685	2.77
multilingual E5	1.000	0.947	11,451 / 22,552	2.78
mGTE	1.000	0.929	10,021 / 22,165	2.79
Qwen3-0.6B	1.000	0.923	10,186 / 21,699	2.77
Qwen3-4B	1.000	0.912	10,048 / 21,232	2.76

Group structure and coverage

Model	Chained	Interleaved	Relevant and twin in top ten / in the same group
BGE-M3	0.047	0.303	0.888 / 0.807
BGE-large	0.045	0.281	0.877 / 0.816
multilingual E5	0.011	0.307	0.923 / 0.907
mGTE	0.037	0.242	0.903 / 0.846
Qwen3-0.6B	0.027	0.263	0.854 / 0.808
Qwen3-4B	0.047	0.269	0.884 / 0.818

MRR@10 under the policy, controls, and oracles

Model	Tie-break	Random order	Random subsets	Oracle group / top ten
BGE-M3	0.598	0.522	0.495	0.817 / 0.908
BGE-large	0.572	0.509	0.462	0.805 / 0.894
multilingual E5	0.610	0.542	0.489	0.883 / 0.939
mGTE	0.584	0.521	0.504	0.827 / 0.919
Qwen3-0.6B	0.546	0.490	0.443	0.780 / 0.871
Qwen3-4B	0.589	0.518	0.489	0.817 / 0.900

F.2 A GROUPING RULE BASED ON THE POLARITY SPAN

We change the grouping rule by linking candidates whose difference lies sufficiently within the canonical span. The statistic is $\rho_{ij} = \|P(z_i - z_j)\|^2 / \|z_i - z_j\|^2$. A threshold chosen on the train+validation corpus matches the cosine policy's number of reordered pairs. This rule corrects fewer twin orders and creates fewer new errors. Its MRR exceeds raw on four encoders and falls below raw on two, but remains below cosine grouping on all six. It also incurs costs on other tasks. Both rules use stored vectors to decide which documents may exchange positions.

Table 31: The same gate with other scorers, and a geometric gate. Upper panel: MRR@10 and twin precedence on separate lines of the near-duplicate tie-break when the within-component order comes from the canonical projection on both sides, on the query only, the W_{doc} lens, the leading singular direction of NevIR-train document differences, the external mean-difference direction of Appendix E (normalized mean of the synthetic affirmative–negative differences, scale selected on Jina triplets), and a random direction (mean of 10). Lower panel: a gate that links two candidates when the share $\rho_{ij} = \|P(z_i - z_j)\|^2 / \|z_i - z_j\|^2$ of their difference inside the canonical span exceeds θ, with θ chosen on the train+validation corpus to match the cosine gate's number of changed candidate pairs; MRR@10 / twin precedence, corrections / new errors, the AUC with which ρ separates the relevant–twin pair from the other top-10 pairs, and the nDCG@10 change on NegConstraint and SciFact for the cosine gate / the ρ gate.

Scorer within cosine groups: MRR@10 above twin precedence

Model	Canonical both sides	Query only	W_{doc}	NevIR SVD	External mean	Semantic Adapter	Random
BGE-M3	0.598 0.657	0.592 0.650	0.596 0.663	0.577 0.607	0.590 0.643	0.582 0.622	0.579 0.615
BGE-large	0.572 0.621	0.579 0.632	0.582 0.649	0.559 0.597	0.579 0.632	0.560 0.607	0.558 0.600
multilingual E5	0.610 0.609	0.607 0.595	0.621 0.642	0.577 0.564	0.606 0.604	0.573 0.557	0.585 0.568
mGTE	0.584 0.625	0.576 0.602	0.582 0.623	0.553 0.566	0.583 0.623	0.552 0.563	0.554 0.560
Qwen3-0.6B	0.546 0.612	0.548 0.607	0.547 0.624	0.518 0.548	0.549 0.608	0.524 0.557	0.526 0.559
Qwen3-4B	0.589 0.641	0.586 0.626	0.590 0.649	0.564 0.581	0.582 0.627	0.569 0.592	0.566 0.588

Geometric grouping and costs

Model	θ	MRR / twin	Corrections / errors	AUC	NegConstraint cosine / ρ	SciFact cosine / ρ
BGE-M3	0.05	0.580 0.638	193 127	0.758	+0.001 −0.018	+0.000 −0.002
BGE-large	0.03	0.551 0.623	206 173	0.733	−0.000 −0.006	+0.000 −0.006
multilingual E5	0.02	0.567 0.599	287 228	0.704	−0.000 −0.033	+0.000 −0.021
mGTE	0.02	0.562 0.613	286 145	0.718	+0.000 −0.012	+0.000 −0.004
Qwen3-0.6B	0.03	0.532 0.603	276 160	0.738	−0.000 −0.103	+0.000 −0.002
Qwen3-4B	0.02	0.573 0.626	234 127	0.728	+0.000 −0.039	+0.000 −0.002

F.3 RANK SLOTS AND METRIC CHANGES

Using f_{10} from Equation 3, suppose the relevant document's component occupies raw rank slots $r_1 < \cdots < r_\ell$. If its within-component rank changes from j_0 to j_1, its RR@10 change is exactly

$$\Delta RR@10 = f_{10}(r_{j_1}) - f_{10}(r_{j_0}). \tag{21}$$

The component keeps its original rank slots, and its own members fill those slots in the new order. A relevant document already at rank 1 can only stay first or move down. Moving one from rank 2 to rank 1 instead gains $1/2$ in reciprocal rank. Correction and error counts alone do not determine MRR change; each change is weighted by the answer's original and new ranks. On NevIR the relevant document shares a component with documents other than its twin on 0.785 to 0.893 of queries and with its twin alone on only 0.028 to 0.056, because the paraphrase edit of the same passage also passes the 0.98 threshold.

The overall MRR@10 change of $+0.010$ to $+0.029$ combines gains and losses at different initial ranks: relevant documents that were already first can only lose and contribute -0.076 to -0.046 (233–378 of 825–986 such queries lose), those at slot two contribute $+0.035$ to $+0.047$, slots three to five $+0.030$ to $+0.046$ and six to ten $+0.001$ to $+0.003$ (Table 32). Improvements below first place offset losses at the top; neither correction counts nor slot-specific results alone describe the net gain.

For nDCG@10, let $g(y)$ be the gain assigned by the evaluator to relevance grade y, and let $w(r)$

Table 32: Where the tie-break's MRR@10 gain comes from on the NevIR test corpus. Share of queries whose relevant document is outside the raw top-10, a singleton there, in a component with its twin only, or in a component with other documents (in parentheses: that class's contribution to the MRR@10 change), and the contribution of the last class by the relevant document's raw slot (gains/losses in queries). A query's reciprocal rank can change only when its relevant document shares a component.

Relevant-document groups and their contributions

Model	ΔMRR	Outside	Singleton	Twin only	Other documents
BGE-M3	+0.020	0.092	0.067	0.056 (−0.000)	0.785 (+0.020)
BGE-large	+0.010	0.106	0.056	0.036 (−0.001)	0.802 (+0.011)
multilingual E5	+0.019	0.061	0.018	0.028 (+0.000)	0.893 (+0.018)
mGTE	+0.029	0.081	0.047	0.050 (+0.001)	0.822 (+0.028)
Qwen3-0.6B	+0.019	0.129	0.044	0.033 (+0.001)	0.794 (+0.018)
Qwen3-4B	+0.023	0.100	0.057	0.039 (+0.001)	0.804 (+0.022)

Other-document groups by original rank: contribution (gains/losses)

Model	1	2	3–5	6–10
BGE-M3	−0.044 (0/225)	+0.034 (227/109)	+0.029 (218/31)	+0.001 (29/5)
BGE-large	−0.059 (0/299)	+0.036 (244/138)	+0.034 (269/25)	+0.002 (39/12)
multilingual E5	−0.075 (0/369)	+0.046 (305/155)	+0.046 (342/22)	+0.002 (31/11)
mGTE	−0.053 (0/266)	+0.037 (251/136)	+0.041 (310/24)	+0.003 (48/5)
Qwen3-0.6B	−0.062 (0/312)	+0.035 (234/122)	+0.044 (308/32)	+0.001 (42/17)
Qwen3-4B	−0.054 (0/279)	+0.033 (221/114)	+0.041 (295/27)	+0.002 (34/10)

be its rank discount. A component whose documents all have grade y contributes $g(y)\sum_j w(r_j)$ in every order. The ideal DCG is unchanged, so its nDCG contribution is also unchanged. A mixed-grade component changes the metric when reordering changes the discounted gain. Let E be the queries with at least one such component. With the same top-ten baseline order used before and after reranking,

$$\overline{\Delta\mathrm{nDCG}}_{\mathrm{all}} = \frac{|E|}{|Q|}\overline{\Delta\mathrm{nDCG}}_{E}, \tag{22}$$

because the change outside E is zero. When E is empty, the overall change is zero and the conditional mean is undefined. Retrieval forms groups from cosine scores; this analysis then uses relevance grades to identify E. Unjudged documents receive grade zero. The conditional analysis compares each reordered list with the raw order of that policy's own top-ten set. Using the same list and tie handling makes the equal-grade identity exact. On Quora, the mixed-grade query subsets have negative mean changes for every encoder; the BGE-M3 and multilingual E5 intervals are below zero.

Table 33: Where the gate can act on the natural collections. Trigger: share of queries with a near-duplicate component in the raw top-10; homogeneous: triggered queries whose components all carry one relevance grade (unjudged counting as zero), whose nDCG@10 stays constant under within-component reordering; effective: queries with a mixed-grade component, with their number, the nDCG@10 change on them with a query-bootstrap 95% interval, the queries that gain/lose, and the candidate pairs inside mixed-grade components. The raw baseline is the raw order through the gate's own top-10 set, so the change on homogeneous queries is exactly zero.

Collection	Trigger	Homo-geneous	Effective (n)	ΔnDCG@10 [95% CI]	Gain/loss	Mixed pairs
BGE-M3						
NFCorpus	0.115	0.090	0.025 (8)	+0.0000 [+0.0000, +0.0000]	0/0	8
TREC-COVID	0.660	0.380	0.280 (14)	+0.0000 [+0.0000, +0.0000]	0/0	16
FiQA	0.006	0.006	0.000 (0)	–	0/0	0
SciDocs	0.009	0.006	0.003 (3)	+0.0000 [+0.0000, +0.0000]	0/0	5
Quora	0.437	0.391	0.046 (457)	−0.0083 [−0.0153, −0.0012]	29/53	1,741
BGE-large						
NFCorpus	0.111	0.084	0.028 (9)	−0.0011 [−0.0032, +0.0000]	0/1	9
TREC-COVID	0.900	0.560	0.340 (17)	−0.0010 [−0.0039, +0.0019]	3/3	25
FiQA	0.000	0.000	0.000 (0)	–	0/0	0
SciDocs	0.031	0.021	0.010 (10)	+0.0034 [+0.0000, +0.0081]	2/0	12
Quora	0.338	0.309	0.029 (294)	−0.0005 [−0.0122, +0.0112]	32/34	784
multilingual E5						
NFCorpus	0.136	0.118	0.019 (6)	+0.0000 [+0.0000, +0.0000]	0/0	6
TREC-COVID	0.780	0.560	0.220 (11)	+0.0060 [−0.0027, +0.0171]	2/2	16
FiQA	0.003	0.003	0.000 (0)	–	0/0	0
SciDocs	0.009	0.004	0.005 (5)	−0.0144 [−0.0497, +0.0197]	1/2	7
Quora	0.491	0.427	0.063 (633)	−0.0091 [−0.0167, −0.0015]	70/102	3,164
mGTE						
NFCorpus	0.118	0.090	0.028 (9)	−0.0002 [−0.0007, +0.0000]	0/1	10
TREC-COVID	0.660	0.360	0.300 (15)	−0.0059 [−0.0172, +0.0003]	1/4	15
FiQA	0.002	0.002	0.000 (0)	–	0/0	0
SciDocs	0.003	0.001	0.002 (2)	+0.0626 [+0.0000, +0.1252]	1/0	2
Quora	0.381	0.343	0.038 (384)	−0.0019 [−0.0119, +0.0081]	34/36	1,301
Qwen3-0.6B						
NFCorpus	0.090	0.071	0.019 (6)	+0.0000 [+0.0000, +0.0000]	0/0	7
TREC-COVID	0.840	0.580	0.260 (13)	+0.0097 [−0.0013, +0.0296]	2/2	17
FiQA	0.020	0.020	0.000 (0)	–	0/0	0
SciDocs	0.009	0.007	0.002 (2)	+0.0074 [+0.0000, +0.0149]	1/0	4
Quora	0.255	0.235	0.020 (198)	−0.0046 [−0.0191, +0.0100]	26/24	455

Continued on the next page

Table 33 (continued)

Collection	Trigger	Homo-geneous	Effective (n)	ΔnDCG@10 [95% CI]	Gain/loss	Mixed pairs
Qwen3-4B						
NFCorpus	0.118	0.090	0.028 (9)	$+0.0000$ $[+0.0000, +0.0000]$	0/0	11
TREC-COVID	0.800	0.480	0.320 (16)	$+0.0052$ $[+0.0000, +0.0156]$	2/0	21
FiQA	0.012	0.012	0.000 (0)	–	0/0	0
SciDocs	0.027	0.017	0.010 (10)	$+0.0042$ $[-0.0204, +0.0361]$	1/2	10
Quora	0.263	0.241	0.022 (221)	-0.0092 $[-0.0206, +0.0017]$	17/28	497

G CONTRAST RECOVERY AND READOUT CONTROLS

These experiments measure contrast recovery on cached vectors by varying training labels, rank, training-set size, and scorer family. Retrieval uses all 1,383 NevIR test pairs. The cue probe predicts a cue-count label on its eligible subset, so we report its accuracy separately. The recovery controls test the readout family used in Section 3.2; Appendix B.4 examines the candidates fitted for the selection study.

Table 34: Model-level accessibility and NevIR behavior. Probe intervals and retrieval intervals are pair-clustered 95% bootstrap CIs. p_s and p_d are exact McNemar tests against raw cosine for W_{sent} and W_{doc}, respectively. Raw cosine for Qwen3-4B is 0.199 here and 0.200 in the later analyses, which recompute it from the cached vectors; the two differ on one near-tied pair of 1,383.

Model	Cue probe	Raw cosine	W_{sent}	W_{doc}	(p_s, p_d)
BGE-M3	0.849 [0.809,0.887]	0.267 [0.243,0.290]	0.321 [0.296,0.346]	0.455 [0.429,0.479]	$(2.1\times10^{-5}, 8.1\times10^{-39})$
BGE-large	0.838 [0.797,0.878]	0.256 [0.232,0.279]	0.302 [0.278,0.327]	0.413 [0.387,0.439]	$(2.5\times10^{-4}, 2.3\times10^{-28})$
multilingual E5	0.887 [0.852,0.919]	0.188 [0.167,0.209]	0.273 [0.251,0.296]	0.392 [0.367,0.417]	$(1.3\times10^{-12}, 6.6\times10^{-44})$
mGTE	0.823 [0.786,0.864]	0.148 [0.130,0.166]	0.225 [0.204,0.247]	0.336 [0.312,0.362]	$(2.6\times10^{-13}, 1.8\times10^{-37})$
Qwen3-0.6B	0.817 [0.777,0.861]	0.153 [0.135,0.172]	0.252 [0.228,0.275]	0.383 [0.357,0.409]	$(1.2\times10^{-18}, 8.2\times10^{-47})$
Qwen3-4B	0.814 [0.774,0.858]	0.199 [0.179,0.220]	0.290 [0.266,0.313]	0.435 [0.410,0.461]	$(1.1\times10^{-14}, 1.6\times10^{-49})$

Table 35: Closed-form projection across all six encoders, one estimator. From 288 affirmative/negated synthetic differences we take the leading k right singular vectors and amplify at scale α; (k, α) is selected on external Jina triplets without using NevIR examples. Intervals are pair-clustered 95% bootstrap CIs; p is an exact McNemar test against raw cosine. Every encoder improves.

Model	(k, α)	Raw	Closed form [95% CI]	Δ (p)
BGE-M3	$(2, 4)$	0.267	0.408 $[0.383, 0.433]$	$+0.141$ (6×10^{-24})
BGE-large	$(2, 8)$	0.256	0.371 $[0.346, 0.397]$	$+0.115$ (2×10^{-15})
multilingual E5	$(2, 8)$	0.188	0.305 $[0.281, 0.329]$	$+0.117$ (6×10^{-16})
mGTE	$(1, 8)$	0.148	0.372 $[0.346, 0.398]$	$+0.224$ (9×10^{-52})
Qwen3-0.6B	$(2, 8)$	0.153	0.338 $[0.313, 0.364]$	$+0.184$ (1×10^{-37})
Qwen3-4B	$(2, 8)$	0.199	0.378 $[0.354, 0.403]$	$+0.179$ (1×10^{-33})

Mirroring each document pair balances the cue-probe labels and fixes chance at 0.5. The single shuffled-label checks score 0.443–0.606 across encoders. These individual runs are reported in the

records; the permutation tests below provide the reference distributions used for inference. Both tests flip labels at the pair level, keeping the two mirrored rows together.

The fixed-probe test flips only test labels (B=20,000). Its distribution is centered at 0.500 ± 0.027, and each encoder's observed accuracy is higher (z=11.6–14.3; $p < 5 \times 10^{-5}$, with no permutation reaching the observed value). The refit test instead retrains the probe on shuffled training labels (B=1,000), following Ojala & Garriga (2010). This distribution is wider because each shuffled assignment can select a different direction in the high-dimensional space. The observed accuracy exceeds this reference for every encoder ($p \leq 9 \times 10^{-3}$). Three sentence-level shuffled-label runs give 0.497 ± 0.003.

G.1 Labels, repeated fits, and probe features

We fit a second NevIR lens after randomly swapping which document is relevant within each training pair. Evaluating both lenses on the same test pairs shows how their performance differs when the training relevance labels are randomized. Refitting W_{doc} with three seeds gives test accuracy BGE-M3 0.453 ± 0.004, BGE-large 0.407 ± 0.007, multilingual E5 0.393 ± 0.003, mGTE 0.341 ± 0.008, Qwen3-0.6B 0.359 ± 0.023, Qwen3-4B 0.430 ± 0.013 (mean and sample standard deviation). The largest standard deviation is 0.023, compared with mean headroom of $+0.151$ to $+0.230$.

The PAWS comparison tests a different choice: the features given to the probe. We fit a concatenation probe and one-hidden-layer MLPs on the same balanced pairs to check whether the original order-invariant features limited the measured accuracy.

Table 36: **Label and feature controls.** The first panel reports NevIR pairwise accuracy for raw cosine, a rank-256 true-label lens, and a lens fitted after randomly swapping the relevant document within training pairs. True-label lenses exceed their independently selected shuffled-label controls on all six encoders; several shuffled fits also improve on raw. The second panel compares order-invariant PAWS features with concatenated features $[z_1, z_2, z_1 - z_2, z_1 \odot z_2]$, using a linear probe and one-hidden-layer MLPs. Chance is 0.5 on these reorder-only PAWS pairs.

	NevIR lens capacity			PAWS probe features			
Model	Raw	True lens	Shuffled lens	Linear inv.	Linear concat	MLP inv.	MLP concat
BGE-M3	0.267	0.449	0.267	0.624	0.588	0.685	0.622
BGE-large	0.256	0.401	0.256	0.633	0.508	0.661	0.526
multilingual E5	0.188	0.392	0.186	0.627	0.618	0.679	0.594
mGTE	0.148	0.345	0.170	0.616	0.534	0.629	0.558
Qwen3-0.6B	0.153	0.385	0.222	0.663	0.582	0.701	0.649
Qwen3-4B	0.200	0.420	0.279	0.679	0.574	0.693	0.578

True-label lenses outperform their independently selected shuffled-label controls on every encoder. Some shuffled-label fits also beat raw, so improvement over raw alone does not establish a benefit from true labels. The PAWS rows compare features for the separate word-order probe.

G.2 Rank, training set size, and shuffled labels

We vary lens rank and training-set size in separate sweeps. We also fit 20 shuffled-label lenses per encoder. Paired-bootstrap intervals are reported for NevIR and CondaQA.

The rank sweep finds a large NevIR gain at rank one. The separate data-size sweep uses rank 256 and finds a large gain with 95 training pairs. In the label control, all 20 shuffled fits per encoder score below the corresponding true-label lens. Paired intervals show the uncertainty in the gain over raw.

Role of the recovery controls. These tests ask whether fitted readouts exploit information in the stored vectors and how recovery depends on labels, capacity, and fitting data. They do not show that contrast-only validation selects a corpus-preserving function. That selection question is evaluated separately with the fixed banks in Appendices B–D.

Table 37: Capacity and sample-size controls for the NevIR lens (test pairwise accuracy, chance 0.25). Rank curve: λ selected on validation at each rank. Size curve: rank 256 fitted on a fixed random 10–100% subsample of NevIR train. Shuffled: mean ± sd of test accuracy over 20 lenses fitted, with the true lens's (r, λ), on labels with the relevance of each training pair swapped at random (a fresh permutation per fit). Headroom: true lens minus raw cosine with a paired bootstrap 95% interval over test pairs.

(a) Rank sweep

Model	Raw	1	4	16	64	256
BGE-M3	0.267	0.444	0.451	0.458	0.458	0.449
BGE-large	0.256	0.417	0.414	0.408	0.403	0.401
multilingual E5	0.188	0.422	0.401	0.399	0.384	0.392
mGTE	0.148	0.350	0.349	0.349	0.359	0.345
Qwen3-0.6B	0.153	0.354	0.353	0.350	0.343	0.385
Qwen3-4B	0.200	0.435	0.427	0.437	0.429	0.420

(b) Training-set size at rank 256

Model	10%	25%	50%	100%
BGE-M3	0.448	0.461	0.424	0.449
BGE-large	0.398	0.419	0.390	0.423
multilingual E5	0.380	0.377	0.400	0.390
mGTE	0.337	0.341	0.355	0.333
Qwen3-0.6B	0.340	0.359	0.373	0.343
Qwen3-4B	0.442	0.440	0.427	0.426

(c) Label control and headroom

Model	Shuffled (mean ± sd)	Headroom [95% CI]
BGE-M3	0.267 ± 0.000	$+0.182$ $[+0.154, +0.210]$
BGE-large	0.256 ± 0.000	$+0.145$ $[+0.119, +0.172]$
multilingual E5	0.188 ± 0.000	$+0.204$ $[+0.176, +0.231]$
mGTE	0.148 ± 0.000	$+0.197$ $[+0.171, +0.223]$
Qwen3-0.6B	0.249 ± 0.018	$+0.232$ $[+0.201, +0.264]$
Qwen3-4B	0.286 ± 0.020	$+0.221$ $[+0.190, +0.251]$

G.3 A NONLINEAR SCORER

A one-hidden-layer MLP scores the same pairs with the same loss and splits on NevIR, PAWS, and CondaQA. This checks whether another scorer changes the large improvements on polarity tasks or the small gains on argument-order retrieval.

The MLP scores below the residual lens on NevIR and below raw cosine on PAWS, while improving CondaQA by a similar amount to the lens. The MLP improves CondaQA accuracy on all six encoders and NevIR accuracy on five, but lowers PAWS accuracy on all six.

H WHICH DIRECTIONS SUPPORT THE RECOVERED DECISIONS?

We compare the learned update with an external polarity span, measure its overlap, and test accuracy after removing aligned components. Canonical amplification uses one or two directions; these localization analyses use eight. Below, P_k denotes the top-k in-domain span from NevIR training differences, and P_{ext} the top-eight span from external synthetic differences.

Removing P_k, $k \in \{2, 4, 8, 16, 32\}$, erases 88.8–91.5% of the full-data lens gain. For 20 random subspaces grown to remove the same lens energy as P_8, post-ablation accuracy is 0.452 ± 0.004 (minimum 0.445), versus 0.283 after removing P_8. Across 20 bootstrap re-estimates, normalized overlap with the original P_8 is 0.744 ± 0.065. Thus, the in-domain ablation removes most of the full-data BGE-M3 gain; matched random controls do not.

Table 38: A richer readout family. An MLP scorer $s(a, d) = \mathrm{MLP}([a, d, a \odot d, |a - d|])$ (one hidden layer of 128 or 256 units, selected on validation with the weight decay) is fitted with the lens loss on each property's train split and scored on its test split. NevIR: pairwise accuracy (chance 0.25); PAWS: paraphrase query against a reordered distractor (chance 0.5); CondaQA: paraphrase against negation edit (chance 0.5).

	NevIR			PAWS		CondaQA	
Model	raw	residual lens	MLP	raw	MLP	raw	MLP
BGE-M3	0.267	0.449	0.334	0.838	0.789	0.656	0.785
BGE-large	0.256	0.401	0.291	0.865	0.842	0.651	0.778
multilingual E5	0.188	0.392	0.181	0.782	0.688	0.594	0.757
mGTE	0.148	0.345	0.198	0.782	0.733	0.650	0.765
Qwen3-0.6B	0.153	0.385	0.215	0.857	0.831	0.619	0.775
Qwen3-4B	0.200	0.420	0.249	0.891	0.846	0.685	0.805

Table 39: Full-data BGE-M3 intervention and in-domain localization. Values are NevIR pairwise accuracy. The matched control removes more learned-lens energy than the P_8 ablation.

Readout or control	Pairwise accuracy
Raw cosine	0.267
Learned W_{doc}	0.455
W_{doc}, remove P_8 (31.1% energy)	0.283
W_{doc}, remove matched control (59.8% energy)	0.427
Closed-form projection	0.408
Projection component parallel to P_8	0.403
Projection component orthogonal to P_8	0.255

Table 40: Localization in a polarity subspace constructed without NevIR data. P_{ext} is the top eight-dimensional span of the synthetic affirmative–negative differences. Energy is the fraction of the NevIR-trained W_{doc} update lying in the span.

Model	Energy in P_{ext}	In-domain P_8	Random span
BGE-M3	0.310	0.346	0.008
BGE-large	0.350	0.422	0.006
multilingual E5	0.302	0.388	0.008
mGTE	0.399	0.494	0.011
Qwen3-0.6B	0.284	0.362	0.007
Qwen3-4B	0.290	0.287	0.004

Table 41: Share of the repair gain removed by output-space ablation of P_{ext} from the lens output, against a random span of the same rank.

Model	Gain removed by P_{ext}	Random span
BGE-M3	19.4%	−0.5%
BGE-large	28.5%	+0.5%
multilingual E5	11.7%	−0.4%
mGTE	23.4%	−1.0%
Qwen3-0.6B	13.3%	+0.1%
Qwen3-4B	14.5%	+0.3%

Table 42: Localization in a polarity span built from an independent synthetic inventory. P_{fresh} is the eight-dimensional span of 288 affirmative–negative differences whose subjects, predicates, and templates are all disjoint from the inventory behind the canonical projection and P_{ext}. Energy: fraction of the NevIR-trained W_{doc} update in each span; gain removed: output-space ablation; overlap: energy of P_{fresh} in P_{ext}, and of the canonical axis in P_{fresh}.

	Lens energy			Gain removed			Overlap	
Model	P_{fresh}	P_{ext}	random	P_{fresh}	P_{ext}	random	fresh in ext	axis in fresh
BGE-M3	0.331	0.310	0.008	19.4%	19.4%	−0.5%	0.303	0.760
BGE-large	0.360	0.350	0.006	31.5%	28.5%	+0.5%	0.323	0.730
multilingual E5	0.321	0.302	0.008	12.8%	11.7%	−0.4%	0.233	0.461
mGTE	0.411	0.399	0.011	31.1%	23.4%	−1.0%	0.302	0.888
Qwen3-0.6B	0.331	0.284	0.007	12.6%	13.3%	+0.1%	0.291	0.691
Qwen3-4B	0.269	0.290	0.004	13.0%	14.5%	+0.3%	0.272	0.605

The NevIR lens and synthetic span use different fitting data. To test the dependence on synthetic wording, a second inventory changes the subjects, predicates, and templates (Table 42). Its overlap and ablation results are reported with the same random-span controls.

With this disjoint inventory, update overlap remains above the random-span control, and ablation removes part of the accuracy gain on all six encoders (Table 42).

Table 43: Controls for shared-row alignment, generic amplification, and explicit cue count on all six encoders. Cross-fit estimates P_8 and the lens on disjoint NevIR halves. Its ablation removes output-space components, whereas the full-data in-domain test removes components from the update factors. Random amplification uses 200 orthonormal subspaces at the canonical rank and scale. Columns: cross-fit lens energy in P_8 (matched random subspace in parentheses); cross-fit fraction of lens gain removed by P_8; canonical amplification Δ against the random-amplification Δ (mean $\pm$ sd); equal-cue raw $\rightarrow$ projection accuracy and its bootstrap 95% CI.

Model	Energy (rand.)	Removed	Canon. Δ	Random Δ	Equal-cue	Equal-cue Δ [CI]
BGE-M3	0.29 (0.01)	18%	+0.141	$+0.004 \pm 0.008$	$0.252 \rightarrow 0.384$	+0.132 [0.102, 0.165]
BGE-large	0.29 (0.01)	31%	+0.115	-0.028 ± 0.010	$0.238 \rightarrow 0.355$	+0.118 [0.088, 0.149]
multilingual E5	0.38 (0.01)	17%	+0.117	-0.034 ± 0.010	$0.188 \rightarrow 0.279$	+0.092 [0.061, 0.124]
mGTE	0.42 (0.01)	28%	+0.224	$+0.000 \pm 0.009$	$0.133 \rightarrow 0.365$	+0.232 [0.200, 0.264]
Qwen3-0.6B	0.28 (0.01)	15%	+0.184	-0.006 ± 0.011	$0.150 \rightarrow 0.325$	+0.174 [0.143, 0.204]
Qwen3-4B	0.19 (0.003)	10%	+0.179	-0.004 ± 0.008	$0.191 \rightarrow 0.359$	+0.169 [0.137, 0.200]

The disjoint-half comparison tests alignment when the lens and its reference span use separate examples. The canonical gain exceeds that from random amplification. Gains also remain on equal-cue pairs, where the two documents have the same detected cue count and a count-only rule ties them.

Tense and quantifier directions provide a structured control on six encoders (Table 44). Tense and quantifier axes built with the same recipe are effective on their own property tests (cross-template AUC 0.971 to 1.000 and 0.941 to 1.000 after projection), yet the tense axis changes NevIR by -0.081 to -0.036 (below zero on six encoders) and the quantifier axis by -0.011 to $+0.026$, against $+0.115$ to $+0.224$ for the polarity axis. The tested structured control directions do not reproduce the polarity direction's gain.

I ALTERNATIVE SCORERS AND RELEVANCE TASKS

The cue probes below predict a count-based label; the retrieval scorers choose a relevant document. We keep these targets separate when comparing lexical baselines and alternative readouts.

Table 44: Tense and quantifier axes compared with the polarity axis. We build tense (present/past) and quantifier (all/some) axes from their own synthetic minimal pairs using the polarity construction. Rank and scale are selected on each construction half's cross-template AUC. The selected axes are then applied to NevIR on both sides. Columns report changes in test pairwise accuracy with pair-bootstrap 95% intervals, and each axis's own cross-template AUC on held-out propositions before and after projection.

		Polarity	Tense axis		Quantifier axis	
Model	Raw	Δ	Δ [95% CI]	own AUC	Δ [95% CI]	own AUC
BGE-M3	0.267	+0.141	−0.081 [−0.103, −0.059]	0.013 → 0.988	+0.015 [+0.004, +0.027]	0.443 → 1.000
BGE-large	0.256	+0.115	−0.074 [−0.096, −0.052]	0.026 → 0.983	+0.009 [−0.012, +0.031]	0.219 → 1.000
multilingual E5	0.188	+0.117	−0.061 [−0.082, −0.043]	0.000 → 1.000	−0.002 [−0.015, +0.011]	0.054 → 1.000
mGTE	0.148	+0.224	−0.036 [−0.055, −0.017]	0.020 → 0.982	+0.026 [+0.004, +0.048]	0.404 → 1.000
Qwen3-0.6B	0.153	+0.184	−0.047 [−0.068, −0.027]	0.011 → 0.999	−0.011 [−0.033, +0.014]	0.138 → 0.941
Qwen3-4B	0.200	+0.179	−0.046 [−0.069, −0.025]	0.001 → 0.971	+0.001 [−0.019, +0.020]	0.201 → 0.998

Different labels and a negation–paraphrase reference. On the 345 unequal-cue test pairs, raw BGE-M3 retrieval accuracy is 0.310, compared with 0.267 over all test pairs and 0.252 on equal-cue pairs. The cue probe reaches 0.849 on those 345 items for a different target: which document has more detected cues. Retrieval gain is measured against the relevance labels, while this probe measures access to cue-count information.

The near-duplicate documents have mean cosine 0.991. For isolated synthetic sentences, affirmative–negative cosine is 0.794, compared with 0.923 for affirmative–paraphrase pairs. Cosine separates the two relations in that setting. The recovery experiments instead evaluate relevance ranking between near-copy documents.

Diagonal relevance probe on all pairs. Cosine decides each NevIR query by $\mathrm{sign}(\mathbf{1}^\top x)$ with $x = z(q) \odot \big(z(d_{\mathrm{rel}}) - z(d_{\mathrm{irrel}})\big)$. Fitting the weights of the same interaction features on NevIR-train (C selected on validation; each decision mirrored so chance is 0.5 per decision and 0.25 per pair, the scale of pairwise accuracy) changes test pairwise accuracy by -0.010 to $+0.056$ and is significant at $p < 0.05$ on three of six encoders (below 0.01 on two) (Table 45). Shuffled-label refits and random weight vectors score 0.04–0.08. The coordinate-mixing readouts on the same pairs reach 0.305–0.408 (closed form) and 0.336–0.455 (W_{doc}): the tested coordinate-wise estimator recovers less than readouts that combine coordinates. The cue-count rule scores 1.000 because the label is defined by that count. In the unmirrored comparison on the same 345 test pairs, TF–IDF difference accuracy is 0.881 with cues and 0.713 after deleting them, against a 0.696 majority baseline. The embedding probes score 0.806–0.872. With the main probe's mirrored augmentation, TF–IDF instead scores 0.997 with cues and 0.712 without them, while embedding probes score 0.814–0.887. These are two augmentation settings for the cue-count task; the retrieval tests use separate relevance labels.

Method-comparison details. Semantic Adapter (Cao, 2025) reweights coordinates as $x \mapsto \mathrm{normalize}(x \odot w)$. Following its published rule, we average the per-dimension contribution difference $\bar{v}$ between each triplet's paraphrase and negated member, divide by $\max(\bar{v})$, and set $w = \mathrm{softmax}(a\bar{v})$ on both sides. The search for a covers 0 to 5 in steps of 0.25. We replace its STS-B negation triplets with the external Jina triplets used to choose our rank and scale, keeping the selection data matched. On all six encoders, the largest-magnitude entry of $\bar{v}$ is positive (its maximum ranges from 0.0008 to 0.0033). Dividing by $\max(\bar{v})$ or by $\max|\bar{v}|$ therefore gives identical weights for these recorded vectors. The selected a ranges from 1.75 to 3.50.

The RMS-coordinate variant instead sets $w_j = 1 + \beta$ on the k_{rms} coordinates with the largest RMS affirmative-minus-negation shift and $w_j = 1$ elsewhere. The same selection data give $k_{\mathrm{rms}} = 64$

Table 45: Diagonal relevance probe versus coordinate-mixing readouts on all 1,383 NevIR test pairs (pairwise accuracy, chance 0.25). The probe reweights cosine's own interaction features; p is an exact McNemar test against cosine.

Model	Cosine	Diagonal probe [95% CI]	Δ (p)	Closed form	W_{doc}
BGE-M3	0.267	0.300 [0.276, 0.325]	+0.033 (0.016)	0.408	0.455
BGE-large	0.256	0.246 [0.224, 0.268]	−0.010 (0.47)	0.371	0.413
multilingual E5	0.188	0.182 [0.163, 0.202]	−0.006 (0.68)	0.305	0.392
mGTE	0.148	0.204 [0.183, 0.225]	+0.056 (3×10^{-6})	0.372	0.336
Qwen3-0.6B	0.153	0.178 [0.159, 0.198]	+0.025 (0.060)	0.338	0.383
Qwen3-4B	0.200	0.252 [0.231, 0.275]	+0.053 (9×10^{-5})	0.378	0.435

and $\beta = 8$; Table 46 identifies this as a variant. The published Semantic Adapter rule raises NevIR pairwise accuracy by +0.013 and lowers NegConstraint nDCG@10 by 0.044. The RMS variant has both a larger gain (+0.037) and a larger cost (0.126). At its selected operating point, the two-sided projection scores higher than Semantic Adapter on NevIR (0.408 versus 0.280) and lower on NegConstraint (0.481 versus 0.681). Query-only projection scores 0.386 and 0.702, respectively.

DEO is reported from its released pipeline's cached BGE-M3 output on NevIR: weight configuration $r = 0.2, p = 1, n = 1$, 20 steps, and local Qwen2.5-1.5B decomposition. Its pairwise accuracy changes from 0.266 to 0.269. This row retains that pipeline's own baseline and settings.

Table 46: Matched post-hoc comparison on frozen BGE-M3. The diagonal and low-rank methods select their scale on external Jina triplets; the Semantic Adapter row follows the published rule (softmax weights from mean per-dimension contribution differences rescaled by their maximum, both sides reweighted), the next row is the RMS-selected variant reported in earlier versions; DEO is its released pipeline, read from its cached NevIR output, and its parenthetical change is relative to its own raw baseline (0.266), the others to raw cosine (0.267).

Readout	Application	NevIR	NegConstraint
Raw cosine	none	0.267	0.725
Semantic Adapter (published rule, $a = 2.5$)	both sides	0.280 (+0.013)	0.681
RMS-selected diagonal variant ($m = 64$, $\beta = 8$)	both sides	0.304 (+0.037)	0.599
DEO (released)	query side	0.269 (+0.003)	—
Canonical low-rank projection	both sides	**0.408** (+0.141)	0.481
Canonical low-rank projection	query only	0.386 (+0.119)	0.702

I.1 Changes in the words expressing the contrast

Cue-reworded and long-context negation. On the already observed CondaQA test split, the raw BGE-M3 separation AUC is 0.605 and W_{sent} reaches 0.629. A balanced probe on cue-reworded paraphrases reaches 0.586; the tested MLP reaches 0.509, providing no nonlinear improvement. With the original explicit-cue cases retained, the linear probe reaches 0.614. The paraphrases change the wording of the negation and retain an average of 0.75 regex-detected markers.

Antonym-family transfer. The canonical axis is built from propositions with explicit *not*, *no*, *never*, or *cannot*. The first holdout contains 144 near-duplicate items whose documents express polarity through 12 antonym families such as *safe/dangerous* and *legal/illegal*. Each query retains *not* but includes the adjective of the wrong document, so surface overlap favors the incorrect candidate. Four encoders start below ceiling on this variant and all four improve: BGE-M3 from 0.799 to 0.972 (per-query 0.865 → 0.986), multilingual E5 from 0.576 to 0.660, Qwen3-0.6B from 0.097 to 0.833, and Qwen3-4B from 0.785 to 1.000; for Qwen3-0.6B the trap pulls raw cosine below chance, and the projection reverses it. BGE-large and mGTE are already at 1.000.

A second, fully lexicon-disjoint variant replaces the query-side negation with a second antonym family (*sound/perilous*, *lawful/criminal*, ...), so neither side uses a polarity token from the construction set. Pairwise accuracy is at or near ceiling there (0.931 to 1.000), and we also report the mean correct-minus-incorrect cosine margin: on every encoder the projection raises it from +0.022–+0.142 to

$+0.532$–$+1.237$, whereas 30 matched random subspaces leave it within 0.003 of the raw value (Table 47).

Table 47: Antonym-family transfer. The adversarial variant keeps query-side negation and traps cosine with the distractor's adjective; the fully lexicon-disjoint variant removes negation from the query as well, so no token on either side appears in the construction set. Accuracy is pairwise accuracy over the 144 items; margin is the mean opposite-polarity cosine gap (correct minus incorrect) under raw cosine, under the canonical projection at the encoder's selected scale, and under the mean of 30 matched random subspaces.

Variant	Model	Raw acc.	Proj. acc.	Margin raw	Margin proj.	Margin rand.
Adversarial	BGE-M3	0.799	0.972	$+0.029$	$+0.498$	$+0.030$
	BGE-large	1.000	1.000	$+0.100$	$+1.319$	$+0.099$
	multilingual E5	0.576	0.660	$+0.008$	$+0.440$	$+0.008$
	mGTE	1.000	1.000	$+0.088$	$+1.281$	$+0.091$
	Qwen3-0.6B	0.097	0.833	-0.008	$+0.393$	-0.007
	Qwen3-4B	0.785	1.000	$+0.019$	$+0.741$	$+0.020$
Fully disjoint	BGE-M3	1.000	1.000	$+0.074$	$+0.532$	$+0.075$
	BGE-large	1.000	0.965	$+0.142$	$+1.237$	$+0.141$
	multilingual E5	1.000	1.000	$+0.022$	$+0.569$	$+0.022$
	mGTE	0.931	0.847	$+0.101$	$+1.085$	$+0.104$
	Qwen3-0.6B	1.000	0.986	$+0.047$	$+0.718$	$+0.048$
	Qwen3-4B	0.979	0.986	$+0.064$	$+0.781$	$+0.064$

New items already solved by raw cosine. The fixed BGE-M3 projection was evaluated on minimal pairs with subjects, predicates, and templates disjoint from the construction data. Raw cosine achieved AUC 0.9997, leaving little room to distinguish the projection from the baseline on these items.

Cue-count compatibility baseline. The cue-count compatibility baseline adds $-\beta(c(q) - c(d))^2$ to cosine. It changes NevIR from 0.267 to 0.268, compared with 0.408 for the projection. On the already observed CondaQA test split it changes the separation AUC from 0.605 to 0.606, against 0.675 for the projection. The CondaQA coefficient is tuned on test, and the paraphrases retain detected cues. This test-tuned count score changes the AUC by 0.001.

I.2 Costs on other relevance tasks

The following tests apply the fitted readouts unchanged to other relevance tasks. Table 48 compares strong two-sided amplification with a small query-side update, and Table 49 reports transferred learned lenses. These transfer losses differ from the in-source results in Appendix B.10, but the learned-readout comparisons use separate fits and do not isolate changes in labels or candidate pools.

Ranking loss on cue-triggered queries. A separate cue-gated BGE-M3 lens runs only on queries containing an explicit negation cue. It changes SciFact from 0.641 to 0.594 overall, but on the 53 activated queries its mean nDCG@10 change is -0.268, with bootstrap interval $[-0.376, -0.166]$. This experiment uses a cue-triggered lens, while Appendix F uses the canonical scorer inside cosine groups. Its conditional result measures the cost on queries selected by explicit negation wording.

I.3 A task with little observed headroom

On PAWS (Zhang et al., 2019), a paraphrase query must rank its source above a meaning-changing word reorder. The task-specific training and validation sets contain 1,617 and 243 retrieval items; the test has 266. We fit and select the same lens family, a synthetic role-swap direction, and a shuffled-label lens. Table 51 reports small additional gains for the tested readouts. The archive also includes the broader probe and property-grid results.

Table 48: Contract summary on all six encoders. Raw scores are NevIR pairwise accuracy, NegConstraint nDCG@10, and SciFact nDCG@10; deltas are changes from raw at the encoder's canonical scale applied to both sides, and at $\alpha = 1$ applied to the query side only.

(a) Raw scores: NevIR / NegConstraint / SciFact

Model	α	Raw scores
BGE-M3	4	0.267 / 0.725 / 0.641
BGE-large	8	0.256 / 0.714 / 0.735
multilingual E5	8	0.188 / 0.674 / 0.689
mGTE	8	0.148 / 0.813 / 0.730
Qwen3-0.6B	8	0.153 / 0.754 / 0.683
Qwen3-4B	8	0.199 / 0.770 / 0.738

(b) Changes: NevIR / NegConstraint / SciFact

Model	Both sides canonical α	Query only $\alpha = 1$
BGE-M3	+0.141 / −0.244 / −0.305	+0.059 / −0.004 / −0.015
BGE-large	+0.115 / −0.181 / −0.464	+0.033 / −0.003 / −0.005
multilingual E5	+0.117 / −0.313 / −0.493	+0.015 / −0.004 / +0.003
mGTE	+0.224 / −0.062 / −0.368	+0.038 / −0.000 / +0.000
Qwen3-0.6B	+0.184 / −0.352 / −0.376	+0.030 / +0.001 / −0.003
Qwen3-4B	+0.179 / −0.329 / −0.284	+0.022 / −0.006 / −0.004

Table 49: nDCG@10 after applying the selected lenses to other collections. Scores use cached embeddings and one shared BEIR evaluator. The lenses are refitted with seed 0 for this comparison, and each row uses its own raw baseline. These fits can differ from the separate single-seed fits elsewhere in the third decimal.

Model	Task	Raw	W_{sent}	W_{doc}
BGE-M3	SciFact	0.641	0.424	0.484
BGE-M3	NFCorpus	0.318	0.193	0.267
BGE-M3	ArguAna	0.398	0.301	0.277
BGE-M3	FiQA	0.411	0.275	0.317
BGE-M3	SciDocs	0.164	0.108	0.141
BGE-M3	TREC-COVID	0.535	0.471	0.471
BGE-M3	Quora	0.886	0.863	0.866
mGTE	SciFact	0.730	0.608	0.550
mGTE	NFCorpus	0.365	0.267	0.330
mGTE	FiQA	0.447	0.369	0.389
mGTE	SciDocs	0.182	0.146	0.166
mGTE	TREC-COVID	0.549	0.518	0.544
mGTE	Quora	0.880	0.866	0.865

Table 50: Argument-order reversal on PAWS, reorder-only pairs (502 balanced test pairs, chance 0.5). The probe reads order-invariant pair features; the bigram baseline is the only surface cue that distinguishes two sentences with identical bags of words. Cosine AUC separates meaning-preserving from meaning-changing reorders, before and after the learned lens.

Model	Probe acc. [95% CI]	Shuffled	Bigram baseline	Cosine AUC	AUC after lens
BGE-M3	0.624 [0.582, 0.665]	0.498	0.564	0.672	0.673
BGE-large	0.633 [0.592, 0.673]	0.538	0.564	0.704	0.710
multilingual E5	0.627 [0.584, 0.669]	0.490	0.564	0.569	0.564
mGTE	0.616 [0.570, 0.657]	0.546	0.564	0.693	0.698
Qwen3-0.6B	0.663 [0.622, 0.703]	0.504	0.564	0.704	0.706
Qwen3-4B	0.679 [0.639, 0.721]	0.522	0.564	0.730	0.732

Table 51: Argument-order retrieval contract (266 test items, chance 0.5): a paraphrase query must prefer the source sentence over its meaning-changing reorder. Lens: rank-selected residual lens fitted on PAWS train; shuffled: the same lens fitted with relevant and distractor swapped at random; closed form: synthetic role-swap axis with the matched random-subspace mean in parentheses; diagonal: per-dimension relevance probe. Headroom is lens minus raw.

Model	Raw	Lens	Lens (doc)	Shuffled	Closed (null)	Diagonal	Headroom
BGE-M3	0.838	0.842	0.835	0.838	0.835 (0.839)	0.733	+0.004
BGE-large	0.865	0.865	0.880	0.868	0.868 (0.867)	0.805	+0.000
multilingual E5	0.782	0.778	0.763	0.782	0.759 (0.754)	0.632	−0.004
mGTE	0.782	0.789	0.786	0.786	0.789 (0.786)	0.688	+0.008
Qwen3-0.6B	0.857	0.857	0.861	0.853	0.850 (0.853)	0.816	+0.000
Qwen3-4B	0.891	0.891	0.887	0.895	0.880 (0.890)	0.823	+0.000

J Shared sources and transfer

J.1 Source overlap between NevIR and CondaQA

NevIR was built from CondaQA passages, with matching passages concentrated in corresponding splits (Table 52): 0.991 to 0.993 of the CondaQA train, validation and test passages match a NevIR document of the same-name split, exactly or by 8-gram containment, and none matches only another split. Every one of the 585 train and validation groups scored after the freeze carries a passage that occurs among the NevIR train and validation documents, and every one of the 688 test groups a passage that occurs among the NevIR test documents.

The lens-transfer audit separately checks matches across the fit-to-test boundary: three of the 688 CondaQA test groups contain a passage that occurs among the NevIR train or validation documents, and seven of the 1,383 NevIR test pairs contain a document that occurs among the CondaQA train or validation passages. The cross-dataset results measure transfer between two tasks built on the same passage source. Most target units have no detected match among the other task's fitting and selection texts. The external projection uses neither dataset for construction or selection.

For a fixed scorer, let $d_i \in \{-1, 0, 1\}$ be the change in correctness of evaluation unit i from raw to the transferred lens. Write N for the number of units, k for the number matched by the source audit, and $\Delta = N^{-1} \sum_i d_i$ for the observed gain. Even if every matched unit contributes $+1$, the gain on the remaining units is at least $\max\{-1, (N\Delta - k)/(N - k)\}$. With $k/N = 3/688$ for the NevIR lens on CondaQA and $7/1383$ for the reverse transfer, this lower bound is positive in all 12 encoder–direction comparisons and is at least $+0.061$ (Table 53). The bound leaves a positive gain on the unmatched units in every comparison. It applies to the recorded scorers; a model retrained on different passages would require a new evaluation.

Table 52: Source overlap between CondaQA and NevIR. For each CondaQA partition, the unique passages (original, paraphrase and negation-flipped edits) and the usable groups, the passages that match a NevIR document of the same-name split exactly (after lower-casing and collapsing punctuation and whitespace) or nearly (word 8-gram containment at least 0.5, not an exact match), the share matched in the same-name split, and the share matched only in another split.

CondaQA partition	Passages	Groups	Exact	Near	Same split	Other splits only
train	1,421	474	947	464	0.993	0.000
validation	340	111	228	109	0.991	0.000
test	2,082	688	1,389	677	0.992	0.000

J.2 Paraphrase and negation edits in CondaQA

CondaQA (Ravichander et al., 2022) groups a Wikipedia passage x_0 with a paraphrase x_1 and a negation-flipping edit x_3. For a scorer h, we evaluate whether the paraphrase scores above the negation-flipped edit: $\Pr[s_h(x_0, x_1) > s_h(x_0, x_3)]$, with chance 0.5; raw cosine is the case $h = h_0$. This is the comparison used in the first frozen-projection check in Appendix A.

Table 53: The contribution of source-matched units to the recorded transfer gain. "Observed" is the transferred lens's accuracy gain over raw (Table 55). "Bound" is the minimum gain on the remaining units when every matched unit contributes $+1$: 3 of 688 CondaQA test groups and 7 of 1,383 NevIR test pairs. This algebraic bound uses the recorded counts and keeps each scorer fixed.

	NevIR lens → CondaQA		CondaQA lens → NevIR	
Model	Observed	Bound	Observed	Bound
BGE-M3	$+0.089$	$+0.085$	$+0.070$	$+0.065$
BGE-large	$+0.065$	$+0.061$	$+0.163$	$+0.159$
multilingual E5	$+0.129$	$+0.126$	$+0.150$	$+0.145$
mGTE	$+0.076$	$+0.072$	$+0.228$	$+0.224$
Qwen3-0.6B	$+0.094$	$+0.091$	$+0.097$	$+0.092$
Qwen3-4B	$+0.103$	$+0.099$	$+0.115$	$+0.110$

The symmetric difference probe is fitted on $z(x_0) - z(x_3)$ using the train and validation groups. The canonical projection uses external synthetic differences for construction and Jina triplets for selection. The in-domain lens instead uses CondaQA train triplets (x_0, x_1, x_3), validation selection, and a shuffled-label control. The first train-and-validation check used the fixed projection before any CondaQA lens fit; the lens and five additional encoder analyses came later as exploratory extensions.

Table 54: The same chain on CondaQA (688 test groups). Raw: $\Pr[\cos(e_0, e_1) > \cos(e_0, e_3)]$ for a passage e_0, its paraphrase edit e_1, and its negation-flipping edit e_3 (chance 0.5), with the AUC over the two similarity populations; probe: symmetric difference probe on $z(e_0) - z(e_3)$ with a shuffled-label control; closed form: the encoder's canonical projection, built and selected without NevIR or CondaQA examples; lens: residual lens fitted on CondaQA train and selected on validation, with a shuffled-label lens as control. Headroom is lens minus raw.

Model	Raw	Raw AUC	Probe	Shuffled	Closed form	Lens	Shuffled lens	Headroom
BGE-M3	0.656	0.605	0.860	0.244	0.737	0.786	0.708	$+0.131$
BGE-large	0.651	0.595	0.811	0.375	0.735	0.753	0.624	$+0.102$
multilingual E5	0.594	0.569	0.874	0.227	0.722	0.776	0.594	$+0.182$
mGTE	0.650	0.595	0.818	0.417	0.730	0.740	0.696	$+0.090$
Qwen3-0.6B	0.619	0.568	0.844	0.291	0.735	0.778	0.679	$+0.158$
Qwen3-4B	0.685	0.606	0.866	0.285	0.775	0.836	0.747	$+0.151$

Both the external projection and the CondaQA-trained lens improve this decision on every encoder.

Table 55: Lens transfer between the two tasks. A lens fitted on one dataset's training split and selected on its validation split is applied unchanged to the other dataset's test split; the two datasets are built from one source (Table 52). NevIR: pairwise accuracy (chance 0.25); CondaQA: $\Pr[s_h(e_0, e_1) > s_h(e_0, e_3)]$ (chance 0.5).

	NevIR test			CondaQA test		
Model	Raw	NevIR lens	CondaQA lens	Raw	CondaQA lens	NevIR lens
BGE-M3	0.267	0.449	0.337	0.656	0.786	0.744
BGE-large	0.256	0.401	0.419	0.651	0.754	0.717
multilingual E5	0.188	0.392	0.338	0.594	0.776	0.724
mGTE	0.148	0.345	0.375	0.650	0.744	0.725
Qwen3-0.6B	0.153	0.385	0.250	0.619	0.776	0.714
Qwen3-4B	0.200	0.420	0.315	0.685	0.844	0.788

Lens transfer across the two tasks. For Table 55, the lens is fitted and selected on one task and applied unchanged to the other task's test split. Tables 52 and 53 describe the shared passage families and the contribution of detected cross-split matches.

The NevIR-trained lens improves CondaQA by $+0.065$–$+0.129$ across the six encoders, compared with $+0.094$–$+0.182$ for the CondaQA-trained refit. In the reverse direction, the CondaQA-trained

lens improves NevIR by $+0.070$–$+0.228$, compared with $+0.145$–$+0.232$ for the NevIR refit. Transfer improves both target tasks on every encoder. The in-domain lens remains higher on all six encoders for CondaQA; on NevIR, the transferred lens is higher for BGE-large and mGTE and lower for the other four.

An axis from natural edits. The canonical projection is built from template stimuli. The same construction and selection rule are applied to CondaQA's human-written negation edits (train groups only), and the resulting axis is scored on NevIR and on the other contracts.

Table 56: A polarity axis built from natural negation edits (CondaQA train $e_0 - e_3$ differences) with the canonical selection rule (k, α on Jina triplet AUC), scored where the synthetic axis is scored. Overlap: energy of the natural basis inside the synthetic span. NevIR: test pairwise accuracy; NegConstraint and SciFact: nDCG@10 raw / natural axis on both sides; corpus: NevIR corpus MRR@10 raw / natural axis.

Model	k/α	Overlap	NevIR raw	natural	synthetic	NegConstraint	SciFact	corpus MRR
BGE-M3	2/4	0.530	0.267	0.386	0.408	0.725 / 0.296	0.641 / 0.357	0.577 / 0.225
BGE-large	2/4	0.417	0.256	0.402	0.371	0.714 / 0.503	0.735 / 0.506	0.561 / 0.288
multilingual E5	2/8	0.332	0.188	0.363	0.305	0.674 / 0.246	0.689 / 0.178	0.591 / 0.196
mGTE	2/4	0.382	0.148	0.352	0.372	0.813 / 0.592	0.730 / 0.427	0.556 / 0.416
Qwen3-0.6B	2/4	0.390	0.153	0.267	0.338	0.754 / 0.385	0.683 / 0.550	0.528 / 0.275
Qwen3-4B	2/4	0.305	0.200	0.295	0.378	0.770 / 0.330	0.739 / 0.634	0.567 / 0.314

Rebuilding the axis from CondaQA's human-written edits gives the six-encoder results in Table 56. The natural axis raises NevIR pairwise accuracy by $+0.095$ to $+0.205$ against $+0.115$ to $+0.224$ for the synthetic axis, keeps 0.305 to 0.530 of its energy inside the synthetic span, and has costs on the other contracts (NegConstraint -0.440 to -0.211, SciFact -0.511 to -0.105). Natural edits thus provide another construction of useful polarity-related directions, alongside the observed retrieval costs.

K Computation and reproducibility

Hardware and software. Most analyses ran on one laptop (Apple M4 Pro, 12 cores, 24 GB unified memory) with PyTorch 2.9.1: Metal handled encoder forward passes, and the CPU handled lens fitting and evaluation. One NVIDIA RTX Pro 6000 ran the cross-device replication, the five-collection local-policy analysis in Table 33, the Qwen3-4B antonym holdout in float16 and float32, and the candidate-expansion, preserving-lens, effective-cell, seed-repeat, and selection records. The actual-list and NFCorpus studies use the GPU results stored in their own ledgers (Appendices C and D). These runs use the same analysis code. Python 3.12.0, transformers 4.57.3, sentence-transformers 5.2.3, datasets 4.5.0, numpy 2.4.0, scikit-learn 1.6.1, and scipy 1.17.0 are pinned in the archive's dependency lock file. The Qwen3-Embedding-4B and Qwen3-4B-Base checkpoints use float16 and batches of 2–8 for the laptop runs; the other encoders use float32. Re-running the twelve front- and middle-placement dilution sweeps on an NVIDIA RTX Pro 6000 (CUDA, PyTorch 2.12) reproduced every reported point estimate to within 0.003 of the laptop values across 432 comparisons. Hub revision hashes of every checkpoint and dataset are recorded in a revision manifest that every model and dataset loader reads.

Wall-clock. Encoding takes most of the time. Per encoder, the NevIR, Jina, and SNLI texts (79k) take 5–15 minutes for the four BERT-style encoders and about an hour for Qwen3-4B. Collection, context, layer, and property-map encoding runs take minutes to about two hours each. Cached-vector analyses take seconds to 20 minutes per encoder. The original laptop pipeline took about 30 hours; later GPU preservation and selection runs have separate records. Seeds are 0 unless stated, and each record names its seed.

Reproducing the evidence. The archive includes 115 ledgers, code, and table and figure generators. With the float16 BGE-M3 cache and frozen canonical basis, the CPU script regenerates stimuli, checks the basis hash, and reproduces canonical NevIR accuracy (0.267 to 0.408). Fixed-bank

selection uses separate ledgers. Other embeddings require the public checkpoints at the recorded revisions.

The archive also contains the full context and position grids, base-checkpoint profiles, property map, Korean variants, and additional scale plots. Their descriptions appear in `archive/extended_appendix_before_focus.tex`.