WHEN DOES DENSE RETRIEVAL NEED ASYMMETRIC GEOMETRY? A BIAS–VARIANCE THEORY OF SHARED AND DUAL PROJECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense retrieval powers retrieval-augmented generation, semantic search, and question answering, yet the theoretical basis for choosing between shared and dual query–document projections remains unclear. We introduce a bias–variance theory for low-rank bilinear scoring. Shared projections induce positive-semidefinite operators, whereas dual projections realize arbitrary low-rank operators. We derive their exact approximation gap and prove a local Gaussian boundary: dual has lower risk exactly when squared directional signal exceeds the estimation cost of its additional degrees of freedom. This boundary motivates the Cross-fitted Asymmetry Risk Selector (CARS), which estimates reproducible directional signal from training pairs; its Gaussian counterpart admits exact selection-power and regret formulas. Guided by the theory, we run retrieval experiments across multiple datasets and embedding models. The mean Dual-minus-Shared NDCG@10 advantage more than doubles as query rotation increases from $0°$ to $90°$. In the rank–sample-size grids, Shared wins 13 of 16 cells at $n = 32$, whereas Dual wins all 32 cells at $n = 1024$ and 2048. Consistent with this shift, all 168 comparable operator-risk curves move toward Dual as training data grow. Compared to the two fixed-geometry baselines, CARS reduces held-out regret by 49–96% and achieves 90.1% mean geometry-selection accuracy.

1 INTRODUCTION

Dense retrieval supplies evidence to question-answering and retrieval-augmented generation systems and powers semantic search (Karpukhin et al., 2020; Lewis et al., 2020; Thakur et al., 2021). When adapting frozen query and document embeddings in $\mathbb{R}^p$, one can apply the *same* rank-r projection to both (*shared*) or fit two projections (*dual*), where $1 \le r \le p$. Throughout, "shared" applies the same projection to queries and documents, while "dual" uses separately parameterized projections to model directional differences between them. Their finite-sample tradeoff depends on signal strength and estimation noise. Figure 1 illustrates the two paradigms.

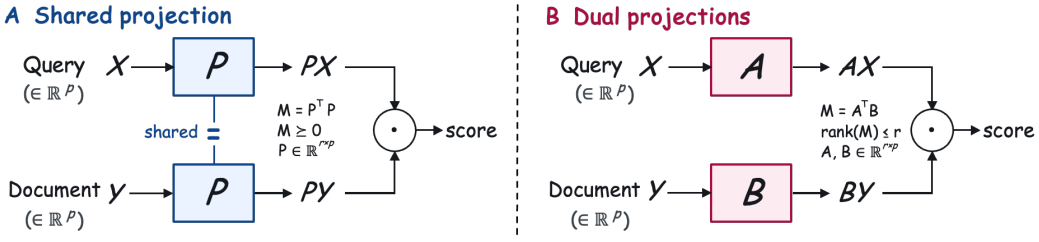

Figure 1: Shared and dual projections for dense retrieval. A shared map P induces a PSD operator $M = P^\top P$, while dual maps A and B induce a general operator $M = A^\top B$ of rank at most r.

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.

The key distinction is geometric. With inner-product scoring, a shared projection $P \in \mathbb{R}^{r\times p}$ gives the similarity matrix $M = P^\top P$, which is positive semidefinite (PSD) because $v^\top M v = \|Pv\|_2^2 \geq 0$ for every v. Separate query and document projections $A, B \in \mathbb{R}^{r\times p}$ instead give $M = A^\top B$, which can represent any rank-at-most-r matrix and need not be symmetric. This flexibility reduces approximation bias but opens more directions to training noise. We ask *when directional signal pays for those extra directions*, and whether the answer can be estimated from training data.

Our contributions are summarized as follows. (i) We characterize the operator classes induced by shared and dual rank-r projections and derive the exact approximation loss imposed by shared projections. (ii) We establish a local Gaussian bias–variance boundary that quantifies when directional signal outweighs the additional estimation cost of dual projections. (iii) We derive a Stein-unbiased rule with exact selection power and regret in the local Gaussian model, then introduce CARS as a cross-fitted selector for real embeddings. (iv) We examine the predicted signal and sample-size effects through controlled simulations, full-corpus retrieval, and held-out operator-risk experiments on real embeddings and datasets.

Related Work. Shared PSD metric learning and unconstrained bilinear similarity learning study the two scoring families (Weinberger & Saul, 2009; Davis et al., 2007; Chechik et al., 2010). Dense-retrieval work compares shared and asymmetric dual encoders and develops methods for adapting pretrained embeddings (Dong et al., 2022; Yoon et al., 2024a; Maekawa et al., 2026). CCA, low-rank geometry, and Stein risk estimation provide related mathematical tools (Hotelling, 1936; Bach & Jordan, 2005; Edelman et al., 1998; Candes et al., 2013; Stein, 1981). What remains unresolved is a theoretical criterion for when the flexibility of dual projections justifies their additional estimation error. We establish an exact boundary in the local Gaussian operator-risk model: Dual has lower risk precisely when its squared directional signal exceeds the variance cost of its additional degrees of freedom. We then develop risk-based selection rules and examine this tradeoff in retrieval and held-out operator-risk experiments.

2 Operator Geometry and Approximation

Let $X, Y \in \mathbb{R}^p$ denote frozen query and document embeddings, and let Y^+, Y^- be a positive and a negative document for query X. With $D = Y^+ - Y^-$, the bilinear margin associated with $M \in \mathbb{R}^{p\times p}$ is

$$m_M(X, Y^+, Y^-) = X^\top M D = \langle M, XD^\top \rangle_F.$$

Here $\langle \cdot, \cdot \rangle_F$ and $\| \cdot \|_F$ denote the Frobenius inner product and norm. The two rank-constrained families are

$$\mathcal{M}_s(r) = \{M = M^\top,\ M \succeq 0,\ \mathrm{rank}(M) \leq r\},$$
$$\mathcal{M}_d(r) = \{M : \mathrm{rank}(M) \leq r\}.$$

A shared projection induces $M = P^\top P \in \mathcal{M}_s(r)$; dual projections induce $M = A^\top B \in \mathcal{M}_d(r)$. Conversely, if $M = U\Sigma V^\top$ has rank at most r, padded factors $A = \Sigma^{1/2}U^\top$ and $B = \Sigma^{1/2}V^\top$ satisfy $A^\top B = M$. Thus $\mathcal{M}_s(r) \subsetneq \mathcal{M}_d(r)$.

Let a population target be $M_\star = H + K$, where

$$H = \frac{M_\star + M_\star^\top}{2}, \qquad K = \frac{M_\star - M_\star^\top}{2}.$$

Write $H = U\,\mathrm{diag}(\eta_1, \ldots, \eta_p)U^\top$ with $\eta_1 \geq \cdots \geq \eta_p$, and let J_r index the largest at most r positive eigenvalues. We first quantify the approximation error of requiring the scoring operator to be shared.

Theorem 1 (Exact shared approximation)**.** *A Frobenius-nearest member of $\mathcal{M}_s(r)$ is $S_r^+ = U\,\mathrm{diag}(\eta_j \mathbf{1}\{j \in J_r\})U^\top$, and the minimum squared error is*

$$\|K\|_F^2 + \sum_{\eta_j < 0} \eta_j^2 + \sum_{\eta_j > 0, j \notin J_r} \eta_j^2. \tag{1}$$

The minimum over $\mathcal{M}_d(r)$ is $\sum_{j>r} \sigma_j(M_\star)^2$; the exact Shared-minus-Dual approximation gap is Equation (1) minus this singular-value tail.

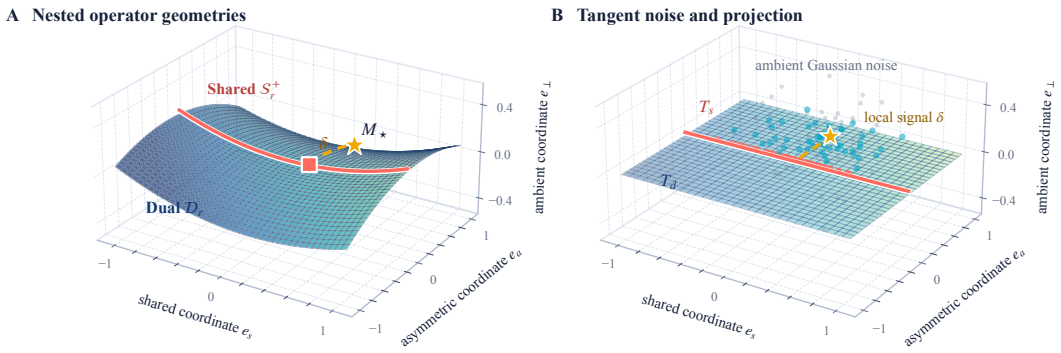

Figure 2: Geometry behind the boundary. **A:** the shared PSD family lies inside the dual family; δ schematically marks a target's departure from sharing. **B:** locally, sharing removes k dual-only noise coordinates but also discards signal along them; k is the extra dimension derived in Section 3.

The three terms in (1) are skew energy, negative eigenvalues, and discarded positive eigenvalues. Appendix A gives the proof. Theorem 1 concerns approximation with a known target. Figure 2A illustrates this nesting: a target outside the shared family can be represented more accurately by dual projections.

Retrieval interpretation. For centered query–positive-document pairs (X, Y^+), let Σ_x, Σ_y be their covariances, and let Y^- be an independent centered document with covariance Σ_y. The relevance moment is $C = \mathbb{E}[X(Y^+ - Y^-)^\top] = \mathbb{E}[XY^{+\top}]$. After whitening the two views, take $M_\star = T = \Sigma_x^{-1/2} C \Sigma_y^{-1/2}$ in Theorem 1. Its Shared-minus-Dual approximation gap then equals the gap in squared optimal rank-r positive–negative separation under a unit negative-score second-moment constraint. Appendix D proves this equivalence and gives a rotation example in which the population advantage of Dual grows with query–document mismatch. With finitely many training pairs, however, Dual can also fit noise in its extra directions. We next ask whether a distribution-free bound quantifies that estimation cost.

Proposition 1 (Global complexity bound)**.** *For n training triples, let $D_i = Y_i^+ - Y_i^-$ and $W_i = X_i D_i^\top$. Suppose $\|W_i\|_F \le R$, and restrict $\|M\|_F \le B$. For $j \in \{s, d\}$, define the empirical Rademacher complexity of the corresponding linear-score class as*

$$\mathfrak{R}_{n,j} = \mathbb{E}_\epsilon \sup_{\substack{M \in \mathcal{M}_j(r) \\ \|M\|_F \le B}} \langle M, \frac{1}{n} \sum_{i=1}^n \epsilon_i W_i \rangle_F.$$

The ϵ_i are independent uniform signs in $\{-1, +1\}$. If $Q = n^{-1} \sum_i \epsilon_i W_i$ and Q_r is its rank-r SVD truncation, then $\mathfrak{R}_{n,s} \le \mathfrak{R}_{n,d} = B\mathbb{E}_\epsilon \|Q_r\|_F \le BR/\sqrt{n}$.

Proposition 1 preserves the ordering $\mathfrak{R}_{n,s} \le \mathfrak{R}_{n,d}$, but bounds both by the same coarse ceiling $BR/\sqrt{n}$. This ceiling does not quantify the extra estimation cost of dual's additional directions. The local noise model in Section 3 resolves that difference; Appendix E gives the proof and a Lipschitz-loss extension.

3 The Bias–Variance Boundary

To measure that cost, we consider a local Gaussian experiment around a rank-r shared operator: an observed population operator is perturbed by isotropic noise of scale $\sigma/\sqrt{n}$, while directional departure from the shared family is also of order $n^{-1/2}$. This common scale makes approximation gain and estimation cost directly comparable.

Define the regular rank-r manifolds

$$\mathcal{D}_r = \{M : \mathrm{rank}(M) = r\}, \qquad \mathcal{S}_r^+ = \{S = S^\top \succeq 0 : \mathrm{rank}(S) = r\}.$$

Let $S = U\Lambda U^\top \in \mathcal{S}_r^+$, where the columns of $U \in \mathbb{R}^{p\times r}$ are orthonormal and $\Lambda \in \mathbb{R}^{r\times r}$ is positive definite. Let T_d and T_s denote the tangent spaces to $\mathcal{D}_r$ and $\mathcal{S}_r^+$ at S, respectively, and let Π_T denote Frobenius-orthogonal projection onto T. Assume that $M_n \in \mathcal{D}_r$ is a sequence of local alternatives with $M_n = S + n^{-1/2}H + O(n^{-1})$ for fixed $H \in T_d$. We observe

$$Y_n = M_n + \frac{\sigma}{\sqrt{n}} G_n,$$

where $\sigma > 0$ and $G_n \in \mathbb{R}^{p\times p}$ has independent standard Gaussian entries. In retrieval, Y_n is a local Gaussian model for the empirical relevance moment $\widehat{C}_n = n^{-1}\sum_{i=1}^n X_i D_i^\top$, while M_n represents its rank-r population target. Let $\widehat{M}_d$ and $\widehat{M}_s$ be Frobenius-nearest points to Y_n in the closed families $\mathcal{M}_d(r)$ and $\mathcal{M}_s(r)$, respectively.

Theorem 2 (Local estimation cost). *In the local model above, $T_s \subset T_d$, $\dim T_d = 2pr - r^2$, and $\dim T_s = pr - r(r-1)/2$. The number of dual-only tangent directions is*

$$k = \dim T_d - \dim T_s = \frac{r(2p - r - 1)}{2}. \tag{2}$$

Moreover, as $n \to \infty$,

$$n\mathbb{E}\|\widehat{M}_d - M_n\|_F^2 \to \sigma^2 \dim T_d,$$
$$n\mathbb{E}\|\widehat{M}_s - M_n\|_F^2 \to \|\Pi_{T_s^\perp} H\|_F^2 + \sigma^2 \dim T_s.$$

Let $\mathcal{A} = T_d \cap T_s^\perp$ denote the dual-only directions, so $\dim \mathcal{A} = k$. The extra dimension decomposes as $k = r(r-1)/2 + r(p-r)$: the first term corresponds to skew perturbations within the active r-dimensional subspace, while the second allows the left and right cross-subspace perturbations to differ.

Corollary 1 (Local phase boundary). *Let $\delta^2 = \|\Pi_{\mathcal{A}} H\|_F^2$. The dual estimator has lower local asymptotic risk than the shared estimator if and only if*

$$\delta^2 > \sigma^2 k. \tag{3}$$

For $\Delta_n = M_n - S = n^{-1/2}H + O(n^{-1})$, $n\|\Pi_{\mathcal{A}}\Delta_n\|_F^2 \to \delta^2$. Away from equality, the same first-order decision compares $n\|\Pi_{\mathcal{A}}\Delta_n\|_F^2$ with $\sigma^2 k$.

The boundary prices the k additional noisy coordinates against the signal they can capture. Figure 2B depicts this local tradeoff: sharing removes variance in $\mathcal{A}$ but also discards the signal there. The synthetic crossing in Figure 3 tests the corresponding signal–sample-size prediction.

3.1 A SELECTION RULE FROM NOISY DATA

Let $Z_n = \sqrt{n}\,\Pi_{T_d}(Y_n - S)$. Since $H \in T_d$, $Z_n = H + \sigma\,\Pi_{T_d} G_n + O(n^{-1/2})$, so Z_n converges in distribution to $Z = H + \sigma G_d$, where G_d is standard Gaussian noise in T_d. We state the selection rule for this limit model, in which the risks of the two projections match Theorem 2.

The boundary in Corollary 1 depends on the unknown δ^2. Its plug-in estimate $\|\Pi_{\mathcal{A}} Z\|_F^2$ is biased upward, since its expectation is $\delta^2 + \sigma^2 k$. We correct this bias with Stein's unbiased risk estimate (SURE) (Stein, 1981), treating σ as known. For $j \in \{s, d\}$, the estimator $\widehat{H}_j = \Pi_{T_j} Z$ has risk $R_j(H) = \mathbb{E}\|\widehat{H}_j - H\|_F^2$, and

$$\widehat{R}_j = \|Z - \Pi_{T_j} Z\|_F^2 + \sigma^2\big(2\dim T_j - \dim T_d\big)$$

is an unbiased estimate of $R_j(H)$. Let $\widehat{\pi} \in \{s, d\}$ select the family with the smaller $\widehat{R}_j$, with ties resolved in favor of s.

Theorem 3 (SURE selection and model-choice regret). *In the model above, $R_s(H) - R_d(H) = \delta^2 - \sigma^2 k$, and $\widehat{\pi} = d$ if and only if*

$$\|\Pi_{\mathcal{A}} Z\|_F^2 > 2\sigma^2 k. \tag{4}$$

Moreover,

$$\Pr(\widehat{\pi} = d) = \Pr\{\chi_k^2(\lambda) > 2k\}, \qquad \lambda = \delta^2/\sigma^2, \tag{5}$$

where $\chi^2_k(\lambda)$ is a noncentral chi-squared variable with k degrees of freedom and noncentrality λ. Define the model-choice regret as $R_{\widehat{\pi}}(H) - \min\{R_s(H), R_d(H)\}$. Its expectation is

$$|\delta^2 - \sigma^2 k| \times \begin{cases} \Pr\{\chi^2_k(\lambda) > 2k\}, & \delta^2 < \sigma^2 k, \\ \Pr\{\chi^2_k(\lambda) \le 2k\}, & \delta^2 > \sigma^2 k, \end{cases} \tag{6}$$

and it is zero when $\delta^2 = \sigma^2 k$.

The threshold $2\sigma^2 k$ combines two terms: one $\sigma^2 k$ removes the expected noise energy in $\mathcal{A}$, and the other is the boundary itself. Thus, unlike the plug-in rule $\|\Pi_{\mathcal{A}} Z\|_F^2 > \sigma^2 k$, SURE rarely selects dual without dual-only signal: when $\delta = 0$, the probability is at most $(2/e)^{k/2}$. Model-choice regret weights the risk gap by the probability of selecting the worse procedure; it is not the post-selection risk of $\widehat{H}_{\widehat{\pi}}$.

3.2 CARS: Cross-fitted Asymmetry Risk Selector

In the local Gaussian experiment, SURE gives an unbiased estimate of the Shared–Dual risk difference when S, T_s, T_d, and σ are specified. For real embeddings, the population geometry and noise covariance are unknown. We therefore use labeled training triples to measure how consistently the dual–shared fit difference appears across disjoint training halves, while penalizing its sampling variation. For a training subset I, define the relevance moment and the dual–shared fit difference

$$\widehat{C}_I = \frac{1}{|I|} \sum_{i \in I} X_i (Y_i^+ - Y_i^-)^\top, \qquad V_I = \mathcal{P}_{d,r}(\widehat{C}_I) - \mathcal{P}_{s,r}(\widehat{C}_I),$$

where $\mathcal{P}_{d,r}$ is rank-r SVD truncation and $\mathcal{P}_{s,r}$ retains the largest r positive eigenvalues of the symmetric part, as in Theorem 1. For each of L random disjoint half-splits $(I_{b,1}, I_{b,2})$ of the training set, the *Cross-fitted Asymmetry Risk Selector* (CARS) computes

$$\widehat{\Gamma}_{\mathrm{CARS}} = \frac{1}{L} \sum_{b=1}^{L} \left[\langle V_{I_{b,1}}, V_{I_{b,2}} \rangle_F - \frac{1}{4} \|V_{I_{b,1}} - V_{I_{b,2}}\|_F^2 \right], \tag{7}$$

and selects Dual if $\widehat{\Gamma}_{\mathrm{CARS}} > 0$, Shared otherwise. The cross-half inner product retains structure that replicates across splits; the disagreement term prices its sampling variation. Under the independent-half model below, the split difference has expected squared norm $4\,\mathrm{tr}(\Sigma)/n$, so the factor $1/4$ charges the full-sample variance cost $\mathrm{tr}(\Sigma)/n$.

This score has a direct connection to the local boundary. Suppose a split residual satisfies $V_{I_{b,j}} = \Delta + \varepsilon_{b,j}$, with independent centered half-sample errors of covariance $2\Sigma/n$. Then

$$\mathbb{E}\widehat{\Gamma}_{\mathrm{CARS}} = \|\Delta\|_F^2 - \frac{\mathrm{tr}(\Sigma)}{n}.$$

If the residual lies in $\mathcal{A}$, $\Delta = n^{-1/2}\Pi_{\mathcal{A}} H$ to first order, and $\Sigma = \sigma^2 I_k$, the right-hand side is $(\delta^2 - \sigma^2 k)/n$. Thus split agreement and disagreement estimate the same signal–variance comparison as Corollary 1, using training data rather than a specified noise variance. The expectation identity is proved in Appendix C.1; the held-out operator-loss study in RQ4 evaluates its geometry choices on real embeddings.

4 Experimental Setting

The theory yields a sequence of empirical questions: whether its signal–uncertainty crossing is visible, whether the signal can be estimated, whether directional mismatch changes retrieval rankings, and whether these estimates improve geometry choice on real embeddings. In real-data Shared–Dual comparisons, both families use paired training subsets and the same frozen embeddings, relevance labels, and test queries. The four base encoders are E5-base-v2, BGE-base, GTE-base, and Contriever-MSMARCO (Wang et al., 2022b; Xiao et al., 2024; Li et al., 2023; Izacard et al., 2022); the real-embedding studies use the subsets specified below. Retrieval quality is measured by mean test-query NDCG@10, defined in Appendix F.

RQ1: Does the predicted boundary appear? We test whether the predicted signal–sample-size crossover appears in controlled two-view retrieval. A Gaussian retrieval experiment rotates the document loading in dimension $p = 32$ and fits ranks $4/8/16$ using 32–2048 training pairs. Each positive is ranked against 299 independent negatives, with five paired seeds at every mismatch–sample-size setting.

RQ2: Can directional signal be estimated? We test whether correcting estimation noise makes observed asymmetry more informative about the held-out benefit of Dual. In the rank-8 simulation, known population PSD distance provides a target for comparing raw plug-in asymmetry with its uncertainty-corrected estimate. On real embeddings, we compare raw and corrected training scores across 840 operator fits on ten tasks. Disjoint held-out relevance moments define the Shared–Dual operator-loss advantage used to assess those scores.

RQ3: How do mismatch, rank, and training size affect retrieval? We test whether directional mismatch changes the relative retrieval quality of Shared and Dual projections. Starting from frozen real embeddings, we rotate only query vectors in eight relevance-informed planes through seven angles from 0° to 90°, leaving document embeddings, evaluation corpora, and relevance labels unchanged. We train rank-16 adapters with 1,024 queries on FEVER, HotpotQA, NQ, MS MARCO, and CQA-TeX using BGE-base and GTE-base. Each angle uses three paired seeds; learning rates and checkpoints are selected on validation queries. To examine the interaction with model capacity and data availability, we also vary rank over $\{4, 8, 16, 32\}$ and training size from 32 to 2,048 queries. Figure 6 shows four complete corpus–encoder grids, with three paired seeds per cell. We rank each complete document collection by exact inner-product search. Figures 5 and 6 report these retrieval comparisons.

RQ4: Does operator risk shift with more data, and can CARS select the better geometry? We test whether more training queries shift held-out operator fit toward Dual and whether CARS can choose the lower-risk geometry using training data alone. The displayed operator-fit curves use NQ and SciFact with all four encoders and three seeds across nested training sizes and ranks. A disjoint query set defines the relevance-moment target M_{test}; for geometry g, normalized loss is $L_g = \|\widehat{M}_g - M_{\text{test}}\|_F^2 / \|M_{\text{test}}\|_F^2$, so positive $L_s - L_d$ favors Dual. Figure 7 displays the seven measured training sizes for NQ (ranks $4/8/16/32$, $n = 256$–2071) and SciFact (ranks $64/128/256$, $n = 256$–647).

For geometry selection, we evaluate CARS (Section 3.2) on five disjoint held-out query folds of ArguAna, MS MARCO, FEVER, NQ, and SciFact, using all four encoders, nested training sizes, ranks $4/8/16/32$, and 20 repeated internal half-splits. For each sample-size–rank cell, selection regret is $L_{\widehat{g}} - \min(L_s, L_d)$; accuracy is the fraction of cells choosing the lower-loss geometry, assigning ties to Shared. Regret is averaged over cells within each held-out fold and then over the five folds. A strict fold win requires lower fold-mean regret than both fixed rules; ties win nothing.

5 Results

5.1 RQ1: Does the predicted boundary appear?

In Figure 3A, the first positive rank-8 grid mean occurs at $n = 256$ for 0.8-radian rotation but at $n = 64$ for 1.2-radian rotation. Figure 3B likewise shows a growing Dual advantage as population distance from the PSD family increases. Together, the panels show the qualitative crossing predicted by Corollary 1: stronger directional mismatch needs fewer training pairs to overcome Dual's estimation cost. The zero contour shows that neither geometry dominates throughout the grid. A separate local-Gaussian calibration matches the risk limits in Theorem 2 within 0.25% and the SURE selection probabilities in Theorem 3 within 0.001 (Appendix Table 2).

5.2 RQ2: Can directional signal be estimated?

The plug-in bias preceding Theorem 3 predicts that a large fitted asymmetry can reflect sampling noise. Figure 4A shows this inflation directly. On real embeddings, the raw plug-in trend is nearly flat across score deciles, whereas the corrected-score trend in Figure 4B rises from Shared-favored to

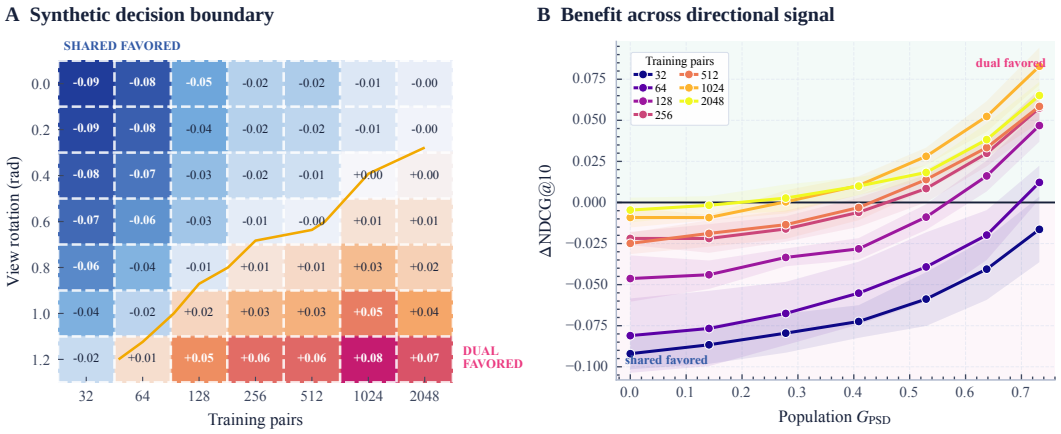

Figure 3: A Shared-to-Dual crossover in controlled two-view retrieval. **A:** Mean Dual-minus-Shared NDCG@10 across mismatch and training size; the gold contour marks zero. **B:** The same gap versus population distance G_{PSD} to the shared PSD family; bands show 95% normal confidence intervals.

Dual-favored held-out fit. Accounting for variation between training splits, as in Equation (7), makes the corrected score more informative about which family has lower operator loss on held-out queries.

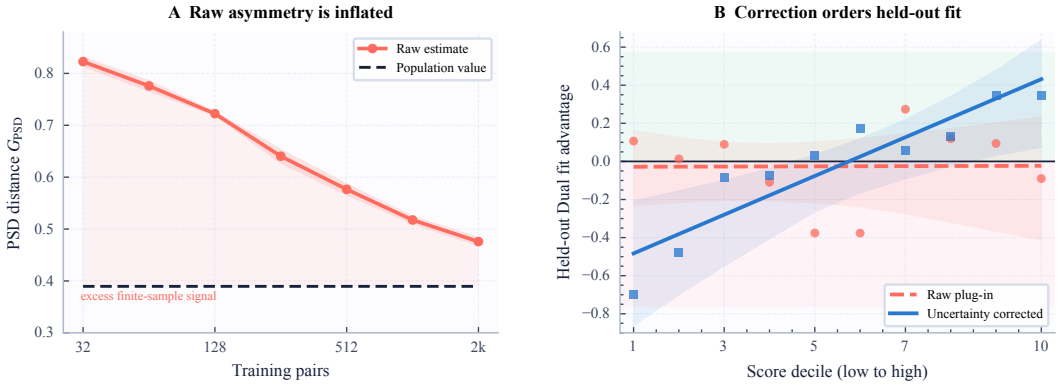

Figure 4: Raw asymmetry is inflated, while correction better predicts held-out fit. **A:** Training PSD distance exceeds its population value; shading shows a 95% t interval. **B:** Held-out Shared-minus-Dual operator-loss advantage across raw and corrected score deciles. Positive values favor Dual.

5.3 RQ3: How do mismatch, rank, and training size affect retrieval?

The rank-one rotation example in Appendix D predicts a widening optimal-separation gap as cross-view mismatch increases. The following experiment asks whether an analogous trend appears in trained, full-corpus retrieval. Figure 5 reports the test NDCG@10 gap at each query rotation. Averaged across the five datasets, the Dual–Shared gap rises from about .0032 at 0° to .0146 at 90°, a 4.6-fold increase. All five dataset means increase between these endpoints; on MS MARCO, the gap changes sign from about $-.0008$ to $+.0219$. The widening gap is consistent with the predicted benefit of Dual projections under stronger cross-view mismatch.

Figure 6 shows a clear effect of training size. Shared wins 13 of the 16 displayed cells at $n = 32$, whereas Dual wins all 32 cells at $n = 1024$ and 2048. Within each grid, rank and rotation are fixed along the sample-size axis, so the initially Shared-favored conditions switch to Dual as training data increase. Averaged over the four displayed grids and seven training sizes, rank 32 yields slightly higher absolute NDCG@10 than rank 4 for both Shared (from 0.7242 to 0.7257) and Dual (from 0.7282 to 0.7289), without enlarging Dual's relative advantage.

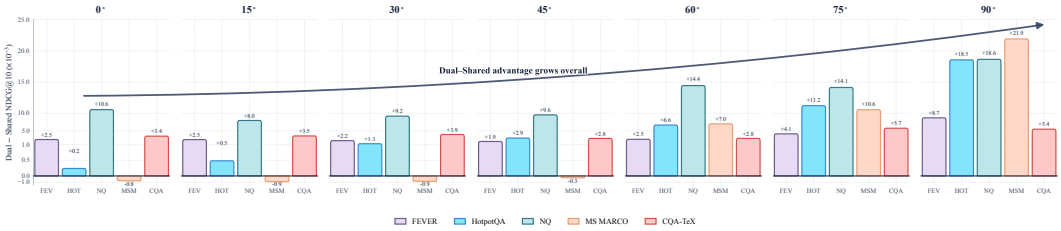

Figure 5: Dual-minus-Shared full-corpus NDCG@10 under query rotation. Each bar averages three paired seeds for BGE-base and GTE-base on each dataset.

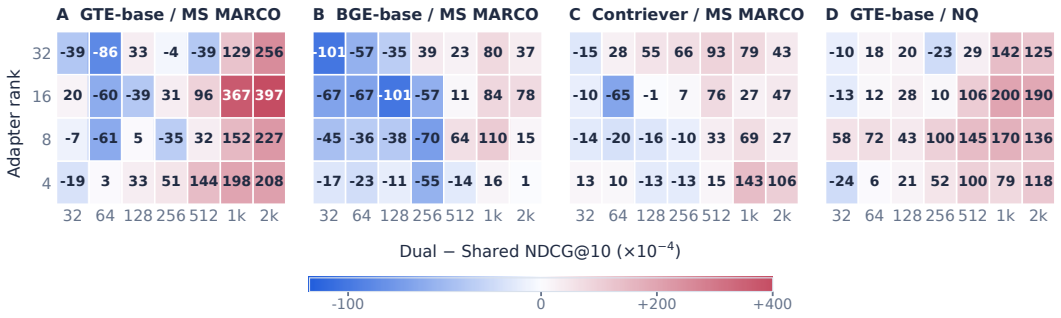

Figure 6: Rank–sample-size retrieval gaps. Each cell is mean Dual-minus-Shared test NDCG@10 over three paired seeds, in units of 10^{-4}; blue favors Shared and red favors Dual.

5.4 RQ4: Does operator risk shift with more data, and can CARS select the better geometry?

The local boundary predicts that, at fixed rank and signal, the variance penalty for Dual diminishes as the number of training queries increases. Across FEVER, NQ, ArguAna, and SciFact, all 168 comparable sample-size slopes are positive (Appendix Figures 8 and 9). Figure 7 shows the encoder means for NQ and SciFact. On NQ, three of the four curves cross from Shared-favored to Dual-favored held-out fit. Together, these results indicate that more data make directional structure easier to estimate.

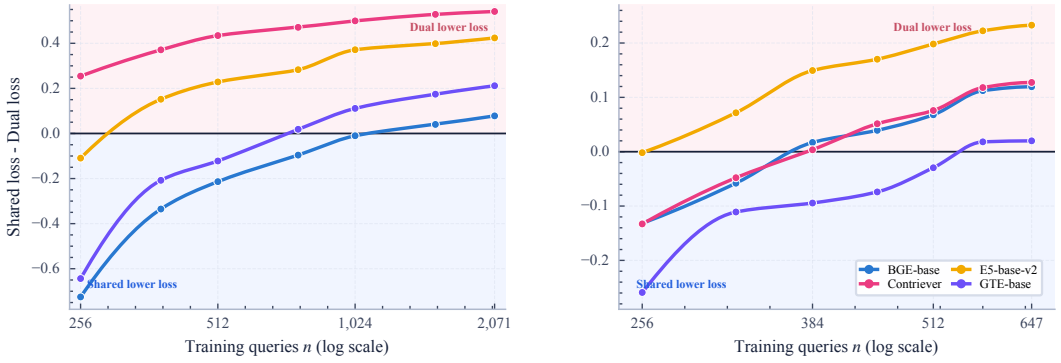

Figure 7: Held-out operator fit moves toward Dual with more training queries. **A:** NQ; **B:** SciFact. Each point averages ranks and three seeds within one encoder; positive values indicate lower held-out loss for Dual.

Geometry selection. We now test the training-only CARS rule in Equation (7) against the fixed Shared and Dual choices. Table 1 shows consistent improvements across all four encoders: CARS wins 85/100 encoder–fold comparisons, reaches 90.1% mean selection accuracy, and reduces mean regret by 49–96% relative to the better fixed choice. This gain answers the geometry-choice part of

RQ4: when the preferred family varies across training sizes, ranks, and datasets, cross-fitted signal minus disagreement is more reliable than committing to one geometry throughout.

Table 1: Geometry selection on five datasets and four encoders. Lower held-out operator-risk regret and higher decision accuracy or strict outer-fold wins indicate better performance. Regret and accuracy are averaged over five held-out query folds.

Rule	Dataset	BGE-base			Contriever			E5-base-v2			GTE-base		
		Regret	Accuracy	Wins	Regret	Accuracy	Wins	Regret	Accuracy	Wins	Regret	Accuracy	Wins
Always Shared	ArguAna	0.0950	45.8%	0/5	0.1310	40.0%	0/5	0.2543	30.0%	0/5	0.1294	43.3%	0/5
	MS MARCO	0.0680	75.6%	1/5	0.3254	46.2%	0/5	0.0839	75.0%	1/5	0.0442	79.4%	1/5
	FEVER	0.2490	33.1%	1/5	0.4959	21.9%	0/5	0.3968	25.0%	0/5	0.2999	37.5%	1/5
	NQ	0.0050	85.0%	1/5	0.1915	40.0%	0/5	0.1020	54.3%	1/5	0.0232	74.3%	2/5
	SciFact	0.0226	63.3%	1/5	0.0288	58.3%	0/5	0.0202	65.0%	2/5	0.0374	65.0%	1/5
Always Dual	ArguAna	0.3342	54.2%	0/5	0.3395	60.0%	0/5	0.1370	70.0%	0/5	0.3310	56.7%	0/5
	MS MARCO	9.5146	24.4%	0/5	2.5724	53.8%	0/5	7.9767	25.0%	0/5	9.6687	20.6%	0/5
	FEVER	0.4271	66.9%	0/5	0.1882	78.1%	0/5	0.2843	75.0%	0/5	0.6001	62.5%	0/5
	NQ	1.5685	15.0%	0/5	0.6073	60.0%	0/5	0.9152	45.7%	0/5	1.5317	25.7%	0/5
	SciFact	0.4445	36.7%	0/5	0.4369	41.7%	0/5	0.5591	35.0%	0/5	0.4899	35.0%	0/5
CARS	ArguAna	**0.0018**	**94.2%**	**5/5**	**0.0059**	**94.2%**	**5/5**	**0.0131**	**88.3%**	**5/5**	**0.0015**	**95.0%**	**5/5**
	MS MARCO	**0.0207**	**93.8%**	**4/5**	**0.0036**	**96.9%**	**5/5**	**0.0251**	**93.1%**	**4/5**	**0.0111**	**91.9%**	**4/5**
	FEVER	**0.1475**	**87.5%**	**4/5**	**0.0749**	**88.1%**	**4/5**	**0.1043**	**86.9%**	**4/5**	**0.2671**	**77.5%**	**4/5**
	NQ	**0.0040**	**90.7%**	**4/5**	**0.0013**	**97.9%**	**5/5**	**0.0568**	**81.4%**	**4/5**	**0.0090**	**92.9%**	**3/5**
	SciFact	**0.0032**	**92.5%**	**4/5**	**0.0042**	**93.3%**	**5/5**	**0.0193**	**86.7%**	**3/5**	**0.0287**	**80.0%**	**4/5**

6 Discussion and future directions

The value of separate projections depends on how query and document representations align after accounting for their marginal covariances. The rotation experiment shows a larger Dual advantage as mismatch increases, while the rank–sample-size grids show that more training data can reverse a low-sample preference for Shared. Geometry choice should therefore account for both directional mismatch and available training data.

Our exact risk boundary assumes local, isotropic Gaussian operator noise. Extending risk estimation to anisotropic noise and directly to ranking metrics would make geometry selection more closely reflect retrieval performance. Another direction is partial sharing, with the number of separately parameterized directions selected alongside rank.

7 Conclusion

We characterized the approximation and estimation costs of shared and dual projections for dense retrieval. In the local Gaussian model, dual projections have lower risk when squared directional signal exceeds $\sigma^2 r(2p - r - 1)/2$. Retrieval and held-out operator experiments show how mismatch, rank, and sample size affect this tradeoff. CARS uses cross-fitted estimates to select between the two projection families, reducing held-out regret by 49–96% relative to the better fixed choice across five datasets. Finally, this framework may guide data-efficient retrieval-head design and other two-view representation problems in which the inputs play different roles.

REPRODUCIBILITY STATEMENT

Detailed derivations and proofs are provided in the appendix, together with experimental protocols. Randomized experiments use predefined seeds for data splits, training, and cross-fitting; results are averaged across multiple seeds where applicable.

AI USE STATEMENT

Generative AI tools were used for language polishing, checking proofs, brainstorming, and reviewing/debugging portions of the experimental code. All AI-generated or AI-modified content has been reviewed and verified by the authors. The authors take full responsibility for all final proofs, code, analyses, and claims.

REFERENCES

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1247–1255, 2013.

Francis R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *ICML*, pp. 89–96, 2005. doi: 10.1145/1102351.1102363.

T. Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018. doi: 10.1214/17-AOS1541.

Emmanuel J Candes, Carlos A Sing-Long, and Joshua D Trzasko. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE transactions on signal processing*, 61 (19):4643–4657, 2013.

Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(36):1109–1135, 2010.

Stéphan Clémençon and Nicolas Vayatis. On partitioning rules for bipartite ranking. In *AISTATS*, volume 5 of *Proceedings of Machine Learning Research*, pp. 97–104, 2009.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *ICML*, pp. 209–216, 2007. doi: 10.1145/1273496.1273523.

Zhe Dong, Jianmo Ni, Dan Bikel, Enrique Alfonseca, Yuan Wang, Chen Qu, and Imed Zitouni. Exploring dual encoder architectures for question answering. In *EMNLP*, pp. 9414–9419, 2022. doi: 10.18653/v1/2022.emnlp-main.640.

Matthias Dorfer, Jan Schlüter, Andreu Vall, Filip Korzeniowski, and Gerhard Widmer. End-to-end cross-modality retrieval with CCA projections and pairwise ranking loss. *International Journal of Multimedia Information Retrieval*, 7(2):117–128, 2018. doi: 10.1007/s13735-018-0151-5.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954.

Chao Gao, Dan Garber, Nathan Srebro, Jialei Wang, and Weiran Wang. Stochastic canonical correlation analysis. *Journal of Machine Learning Research*, 20(167):1–46, 2019.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *ACL*, pp. 1762–1777, 2023. doi: 10.18653/v1/2023.acl-long.99.

Matan Gavish and David L. Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017. doi: 10.1109/TIT.2017.2653801.

Nicholas J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jKN1pXi7b0.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Yizhi Li, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. More robust dense retrieval with contrastive dual learning. In *ICTIR*, pp. 287–296, 2021. doi: 10.1145/3471158.3472245.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

Xueguang Ma, Minghan Li, Kai Sun, Ji Xin, and Jimmy Lin. Simple and effective unsupervised redundancy elimination to compress dense vectors for passage retrieval. In *EMNLP*, pp. 2854–2859, 2021. doi: 10.18653/v1/2021.emnlp-main.227.

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

Rahul Mazumder and Haolei Weng. Computing the degrees of freedom of rank-regularized estimators and cousins. *Electronic Journal of Statistics*, 14(1):1348 – 1385, 2020. doi: 10.1214/20-EJS1681. URL https://doi.org/10.1214/20-EJS1681.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *NAACL*, pp. 5835–5847, 2021. doi: 10.18653/v1/2021.naacl-main.466.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. doi: 10.1214/aos/1176345632.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

Bart Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. doi: 10.1137/110845768.

Bart Vandereycken, P.-A. Absil, and Stefan Vandewalle. A Riemannian geometry with complete geodesics for the set of positive semidefinite matrices of fixed rank. *IMA Journal of Numerical Analysis*, 33(2):481–514, 2013. doi: 10.1093/imanum/drs006.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *NAACL*, pp. 2345–2360, 2022a. doi: 10.18653/v1/2022.naacl-main.168.

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*, 2022b.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(9):207–244, 2009.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 641–649, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *ICLR*, 2021. URL https://openreview.net/forum?id=zeFrfgyZln.

Jinsung Yoon, Yanfei Chen, Sercan Arik, and Tomas Pfister. Search-adaptor: Embedding customization for information retrieval. In *ACL*, pp. 12230–12247, 2024a. doi: 10.18653/v1/2024.acl-long.661.

Jinsung Yoon, Rajarishi Sinha, Sercan O. Arik, and Tomas Pfister. Matryoshka-adaptor: Unsupervised and supervised tuning for smaller embedding dimensions. In *EMNLP*, pp. 10318–10336, 2024b. doi: 10.18653/v1/2024.emnlp-main.576.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. COCO-DR: Combating the distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning. In *EMNLP*, pp. 1462–1479, 2022. doi: 10.18653/v1/2022.emnlp-main.95.

APPENDIX

The appendix provides proofs, protocols for the reported experiments, and the held-out operator-risk trajectories cited in the main text.

A COMPLETE PROOF OF THE APPROXIMATION THEOREM

We first record the strict expressivity relation. For every $P \in \mathbb{R}^{r\times p}$, $P^\top P$ is symmetric, $v^\top P^\top P v = \|Pv\|_2^2 \geq 0$, and $\mathrm{rank}(P^\top P) \leq r$. Conversely, if $M = U\Sigma V^\top$ has rank $q \leq r$, define

$$A = \begin{bmatrix} \Sigma^{1/2} U^\top \\ 0_{(r-q)\times p} \end{bmatrix}, \qquad B = \begin{bmatrix} \Sigma^{1/2} V^\top \\ 0_{(r-q)\times p} \end{bmatrix}.$$

Then $A^\top B = M$. For $1 \leq r \leq p$, the rank-one matrix $-e_1 e_1^\top$ belongs to $\mathcal{M}_d(r)$ but not $\mathcal{M}_s(r)$, so $\mathcal{M}_s(r) \subsetneq \mathcal{M}_d(r)$.

Proof of Theorem 1. For any symmetric S, the Frobenius inner product between the skew-symmetric K and symmetric $H - S$ vanishes. Hence

$$\|M_\star - S\|_F^2 = \|K\|_F^2 + \|H - S\|_F^2.$$

The first term is unavoidable. Diagonalize $H = U\,\mathrm{diag}(\eta)U^\top$. By orthogonal invariance and the Hoffman–Wielandt inequality, an optimal symmetric S can be chosen with the same eigenvectors. PSD constrains its eigenvalues s_j to be nonnegative, and the rank constraint permits at most r nonzero values. A negative η_j is therefore optimally paired with zero. For a positive η_j, retaining it at $s_j = \eta_j$ costs zero and dropping it costs η_j^2. The best support keeps the largest at most r positive eigenvalues. Adding the orthogonal skew residual yields Equation (1).

For comparison with the Dual class, let $M_{\star,r}$ be the rank-r SVD truncation of $M_\star$ and set $a_r^2 = \sum_{j\leq r} \sigma_j(M_\star)^2$. For any D of rank at most r, von Neumann's inequality followed by Cauchy–Schwarz gives $\langle M_\star, D\rangle_F \leq a_r \|D\|_F$. Consequently,

$$\|M_\star - D\|_F^2 \geq \|M_\star\|_F^2 + \|D\|_F^2 - 2a_r\|D\|_F \geq \|M_\star\|_F^2 - a_r^2 = \sum_{j>r} \sigma_j(M_\star)^2.$$

Both inequalities are equalities at $D = M_{\star,r}$. Since $\mathcal{M}_s(r) \subset \mathcal{M}_d(r)$, subtracting this Dual minimum from Equation (1) gives a nonnegative and exact approximation gap between the two families. □

B GEOMETRY AND PROOF OF THE LOCAL RISK THEOREM

B.1 TANGENT SPACES AND DIMENSIONS

For $j \in \{s, d\}$, write $d_j = \dim T_j$. Let $S = U\Lambda U^\top$ with $U \in \mathbb{R}^{p\times r}$ orthonormal and $\Lambda \succ 0$. Extend U to an orthogonal matrix $[U, U_\perp]$.

For the general fixed-rank manifold, first-order perturbation of an SVD gives

$$T_d = \left\{ UAU^\top + U_\perp BU^\top + UCU_\perp^\top : \begin{array}{l} A \in \mathbb{R}^{r\times r}, \\ B \in \mathbb{R}^{(p-r)\times r}, \\ C \in \mathbb{R}^{r\times(p-r)} \end{array} \right\}. \tag{8}$$

The three blocks are orthogonal and free, so $\dim T_d = r^2 + 2r(p-r) = 2pr - r^2$.

For the PSD manifold, use the local representation $S(t) = U(t)\Lambda(t)U(t)^\top$. Differentiating at zero gives

$$T_s = \left\{ UAU^\top + U_\perp BU^\top + UB^\top U_\perp^\top : A = A^\top,\ B \in \mathbb{R}^{(p-r)\times r} \right\}. \tag{9}$$

Therefore

$$\dim T_s = \frac{r(r+1)}{2} + r(p-r) = pr - \frac{r(r-1)}{2}.$$

Every matrix in Equation (9) is a special case of Equation (8), proving $T_s \subset T_d$. Subtraction gives $k = r(2p - r - 1)/2$.

The difference space also has an explicit orthogonal block description. In the basis $[U, U_\perp]$, write $X \in T_d$ as $\left[\begin{smallmatrix} A & C \\ B & 0 \end{smallmatrix}\right]$. Its T_s component has upper-left block $\mathrm{sym}(A)$ and paired cross-blocks $B_s = (B + C^\top)/2$ and $B_s^\top$. The remaining component in $\mathcal{A} = T_d \cap T_s^\perp$ is

$$\begin{bmatrix} \mathrm{skew}(A) & -D^\top \\ D & 0 \end{bmatrix}, \qquad D = \frac{B - C^\top}{2}.$$

The skew block contributes $r(r-1)/2$ coordinates and the independent cross-block D contributes $r(p-r)$, proving the stated decomposition of k. Orthogonality also gives $\|\Pi_{\mathcal{A}} X\|_F^2 = \|\,\mathrm{skew}(A)\|_F^2 + \frac{1}{2}\|B - C^\top\|_F^2$.

The corresponding orthogonal projectors can be written explicitly. With $P = UU^\top$ and $Z_{\mathrm{sym}} = (Z + Z^\top)/2$,

$$\Pi_{T_d}(Z) = PZ + ZP - PZP,$$
$$\Pi_{T_s}(Z) = PZ_{\mathrm{sym}} + Z_{\mathrm{sym}}P - PZ_{\mathrm{sym}}P.$$

B.2 Derivative of nearest-point projection

We give a self-contained first-order argument. Let $\mathcal{N}$ be either regular manifold and choose a smooth local chart $\phi(u) = S + Lu + Q(u)$, where L is injective, $\mathrm{range}(L) = T_S\mathcal{N}$, and $\|Q(u)\| = O(\|u\|^2)$. For small ambient perturbation e, a locally nearest point $\phi(\widehat{u})$ minimizes $\|S + e - \phi(u)\|^2$. Its normal equation is

$$L^\top(e - L\widehat{u}) = O(\|e\|^2),$$

because $\widehat{u} = O(\|e\|)$ and both the chart remainder and derivative remainder are second order. Solving gives

$$L\widehat{u} = L(L^\top L)^{-1}L^\top e + O(\|e\|^2) = \Pi_{T_S\mathcal{N}} e + O(\|e\|^2).$$

Consequently the nearest-point map satisfies

$$\pi_{\mathcal{N}}(S + e) = S + \Pi_{T_S\mathcal{N}} e + O(\|e\|^2). \tag{10}$$

Regularity and $\Lambda \succ 0$ ensure a sufficiently small neighborhood with a locally unique projection. The global nearest points used in the theorem exist because $\mathcal{M}_d(r)$ and $\mathcal{M}_s(r)$ are closed. Since S belongs to both, any nearest point $\widehat{M}_j$ satisfies $\|\widehat{M}_j - S\|_F \le 2\|e\|_F$. For sufficiently small e, the smallest positive singular value of S therefore keeps a nearest point at rank r. It then agrees with the smooth local projection in Equation (10).

Proof of Theorem 2. Set $e_n = n^{-1/2}(H + \sigma G) + O(n^{-1})$ in Equation (10). For $j \in \{s, d\}$,

$$\widehat{M}_j = S + \frac{1}{\sqrt{n}}\Pi_{T_j}(H + \sigma G) + O_p(n^{-1}).$$

Subtracting $M_n = S + n^{-1/2}H + O(n^{-1})$ yields

$$\sqrt{n}(\widehat{M}_j - M_n) = -\Pi_{T_j^\perp} H + \sigma\Pi_{T_j} G + o_p(1).$$

The deterministic bias is normal to the projected Gaussian noise. In any Frobenius-orthonormal basis of T_j, the latter has d_j independent standard normal coordinates, hence expected squared norm d_j. To pass from convergence in probability to risk convergence, use the global nearest-point bound $\|\widehat{M}_j - S\|_F \le 2\|Y_n - S\|_F$. The deterministic remainder in $M_n = S + n^{-1/2}H + O(n^{-1})$ is uniformly bounded after multiplication by $\sqrt{n}$, so, for all sufficiently large n,

$$n\|\widehat{M}_j - M_n\|_F^2 \le C\big(1 + \|H\|_F^2 + \sigma^2\|G\|_F^2\big)$$

for a constant C independent of n. The Gaussian right-hand side is integrable; dominated convergence (after the local expansion, which holds almost surely for each fixed G) therefore gives

$$n\mathbb{E}\|\widehat{M}_j - M_n\|_F^2 \longrightarrow \|\Pi_{T_j^\perp} H\|_F^2 + \sigma^2 d_j.$$

Since $H \in T_d$, the dual bias vanishes. Since $T_s \subset T_d$, the shared bias is its component in $T_d \cap T_s^\perp$. □

Proof of Corollary 1. By the orthogonal decomposition $T_d = T_s \oplus \mathcal{A}$, the shared asymptotic risk exceeds the dual asymptotic risk by

$$\|\Pi_{T_s^\perp} H\|_F^2 + \sigma^2 d_s - \sigma^2 d_d = \|\Pi_{\mathcal{A}} H\|_F^2 - \sigma^2 k = \delta^2 - \sigma^2 k.$$

The difference is positive exactly under Equation (3); equality gives a first-order tie. For $\Delta_n = n^{-1/2} H + O(n^{-1})$, linearity and boundedness of the orthogonal projector give $n\|\Pi_{\mathcal{A}} \Delta_n\|_F^2 = \delta^2 + O(n^{-1/2})$. Hence the unscaled expression is the same first-order boundary along these local alternatives, away from equality. □

Why the theorem is local. At rank-changing points the two sets are stratified rather than smooth, so a tangent cone replaces the tangent space. For a fixed target far outside the PSD manifold, the derivative of its nearest-point projection also contains a curvature (shape-operator) term. The local alternative avoids both issues and is the regime in which effective dimension has an exact first-order meaning.

C Proof of the SURE Selector

Proof of Theorem 3. Work in a Frobenius-orthonormal basis of T_d, so the observation is an ordinary d_d-dimensional Gaussian sequence. For the linear estimator $\widehat{H}_j = \Pi_{T_j} Z$, Stein's formula (Stein, 1981) is

$$\widehat{R}_j = \|(I - \Pi_{T_j}) Z\|^2 + 2\sigma^2 d_j - \sigma^2 d_d.$$

Its unbiasedness is also immediate here without a general Stein identity: $\mathbb{E}\|(I - \Pi_{T_j}) Z\|_F^2 = \|(I - \Pi_{T_j}) H\|_F^2 + \sigma^2 (d_d - d_j)$, so $\mathbb{E}\widehat{R}_j = \|(I - \Pi_{T_j}) H\|_F^2 + \sigma^2 d_j = \mathbb{E}\|\widehat{H}_j - H\|_F^2$. For $j = d$, this reduces to $\sigma^2 d_d$. Because $T_d = T_s \oplus \mathcal{A}$ orthogonally, for $j = s$ it is

$$\widehat{R}_s = \|\Pi_{\mathcal{A}} Z\|^2 + 2\sigma^2 d_s - \sigma^2 d_d.$$

Thus $\widehat{R}_d < \widehat{R}_s$ exactly when $\|\Pi_{\mathcal{A}} Z\|^2 > 2\sigma^2 (d_d - d_s)$, proving Equation (4).

In an orthonormal basis of $\mathcal{A}$, $\Pi_{\mathcal{A}} Z / \sigma$ is a k-variate identity-covariance normal with mean squared norm $\lambda = \delta^2 / \sigma^2$. Its squared norm therefore has the noncentral $\chi_k^2(\lambda)$ distribution, proving Equation (5).

The two unconditional fixed-model risks in the scaled experiment are

$$R_s(H) = \delta^2 + \sigma^2 d_s, \qquad R_d(H) = \sigma^2 d_d.$$

Their difference is $\delta^2 - \sigma^2 k$. Under the model-choice loss defined in Theorem 3, if this difference is negative, regret occurs only when SURE selects dual; if positive, regret occurs only when SURE selects shared. Multiplication by the corresponding exact tail probability proves Equation (6). Replacing σ^2 by σ^2 / n gives the unscaled threshold. □

C.1 Expectation of the cross-fitted CARS score

The real-embedding selector repeatedly splits data into disjoint halves rather than the isotropic Gaussian observation in Theorem 3. Its score has a simple expectation under the stated independent-error model. Write $V_j = \Delta + \varepsilon_j$, where the two errors are independent and mean-zero with covariance $2\Sigma / n$ in an orthonormal operator-coordinate basis. Independence and centering give

$$\mathbb{E}\langle V_1, V_2 \rangle_F = \|\Delta\|_F^2, \qquad \mathbb{E}\|V_1 - V_2\|_F^2 = \mathbb{E}\|\varepsilon_1\|_F^2 + \mathbb{E}\|\varepsilon_2\|_F^2 = \frac{4 \operatorname{tr}(\Sigma)}{n}.$$

Thus

$$\mathbb{E}\left[\langle V_1, V_2 \rangle_F - \tfrac{1}{4}\|V_1 - V_2\|_F^2\right] = \|\Delta\|_F^2 - \frac{\operatorname{tr}(\Sigma)}{n}.$$

Averaging over repeated half-splits leaves this expectation unchanged; independence across the repeated splits is unnecessary for this identity. If the residuals span the k dual-only directions and $\Sigma = \sigma^2 I_k$, then $\operatorname{tr}(\Sigma) = \sigma^2 k$, matching the unscaled first-order boundary in Corollary 1. The identity is an expectation calculation; unlike the Gaussian SURE theorem, it does not assert an exact finite-sample selection probability.

D Optimal Rank-r Separation in Two-View Retrieval

Let (X, Y^+) be a centered query–positive-document pair with positive-definite covariances Σ_x, Σ_y. Let Y^- be a centered document with covariance Σ_y, independent of X. Define the population relevance moment and its whitened form by

$$C = \mathbb{E}[X(Y^+ - Y^-)^\top] = \mathbb{E}[XY^{+\top}], \qquad T = \Sigma_x^{-1/2} C \Sigma_y^{-1/2}.$$

For $j \in \{s, d\}$, the optimal positive–negative separation under a unit negative-score second moment is

$$\mathfrak{D}_j^\star(r) = \sup_{\substack{M \in \mathcal{M}_j(r) \\ \mathbb{E}[(X^\top M Y^-)^2] \le 1}} \mathbb{E}[X^\top M (Y^+ - Y^-)].$$

Independence gives $\mathbb{E}[(X^\top M Y^-)^2] = \|\Sigma_x^{1/2} M \Sigma_y^{1/2}\|_F^2$.

Theorem 4 (Optimal rank-r separation)**.** *Let $\sigma_j(T)$ be the jth largest singular value of T. Then*

$$[\mathfrak{D}_d^\star(r)]^2 = \sum_{j=1}^{r} \sigma_j(T)^2. \tag{11}$$

An optimum is proportional to $\Sigma_x^{-1/2} T_r \Sigma_y^{-1/2}$, where T_r is the rank-r SVD truncation of T. If both views are whitened, so $\Sigma_x = \Sigma_y = I$ and $T = C$, let $\eta_1 \ge \cdots \ge \eta_p$ be the eigenvalues of $\mathrm{sym}(T) = (T + T^\top)/2$ and $\eta_j^+ = \max(\eta_j, 0)$. Then

$$[\mathfrak{D}_s^\star(r)]^2 = \sum_{j=1}^{r} (\eta_j^+)^2. \tag{12}$$

For whitened views, the squared separation gap $[\mathfrak{D}_d^\star(r)]^2 - [\mathfrak{D}_s^\star(r)]^2$ equals the exact Shared-minus-Dual approximation gap in Theorem 1 with $M_\star = T$.

Corollary 2 (When shared geometry is sufficient)**.** *If the views are whitened and T is symmetric PSD, then $\mathfrak{D}_s^\star(r) = \mathfrak{D}_d^\star(r)$ for every r. The same equality holds without preprocessing whitening for matched latent views $X = C_0 Z + \epsilon_q$ and $Y^+ = C_0 Z + \epsilon_d$, where the components are centered and mutually independent, $\mathrm{Cov}(Z) = I$, and $\mathrm{Cov}(\epsilon_q) = \mathrm{Cov}(\epsilon_d)$.*

Proof of Theorem 4. Because Y^- is centered and independent of X,

$$\mathbb{E}[X^\top M (Y^+ - Y^-)] = \mathrm{tr}(M^\top C).$$

Independence also gives

$$\begin{aligned}\mathbb{E}[(X^\top M Y^-)^2] &= \mathbb{E}_X\left[X^\top M \Sigma_y M^\top X\right] \\ &= \mathrm{tr}(M^\top \Sigma_x M \Sigma_y).\end{aligned}$$

Let $N = \Sigma_x^{1/2} M \Sigma_y^{1/2}$. Then rank is preserved, the noise constraint is $\|N\|_F \le 1$, and

$$\mathrm{tr}(M^\top C) = \langle N, T \rangle_F, \qquad T = \Sigma_x^{-1/2} C \Sigma_y^{-1/2}.$$

By von Neumann's trace inequality and Cauchy–Schwarz, for rank at most r,

$$\langle N, T \rangle_F \le \left(\sum_{j \le r} \sigma_j(T)^2 \right)^{1/2} \|N\|_F.$$

If $T_r = 0$, both sides vanish and $N = 0$ is optimal. Otherwise equality is attained by $N = T_r / \|T_r\|_F$, proving Equation (11).

For whitened views and a PSD M, write $T = H + K$ with H symmetric and K skew-symmetric. Then $\langle M, K \rangle_F = 0$. Let the eigenvalues of H be $\eta_1 \ge \cdots \ge \eta_p$. Von Neumann's inequality aligns the eigenvectors of a maximizing PSD M with those of H. The remaining optimization is

$$\max_{m_j \ge 0,\ \|m\|_2 \le 1,\ \|m\|_0 \le r} \sum_j m_j \eta_j,$$

whose value is the Euclidean norm of the largest at most r positive eigenvalues. This proves Equation (12). Since the PSD feasible set is contained in the unrestricted rank-r set, the difference of their squared optima is nonnegative. □

Proof of Corollary 2. When $T = T^\top \succeq 0$, its singular values are exactly its nonnegative eigenvalues, and $\mathrm{sym}(T) = T$. Consequently the spectral sums in Equations (11) and (12) coincide for every r. In the matched latent model, independent view noises imply $C = \mathbb{E}[XY^{+\top}] = C_0 C_0^\top$ and $\Sigma_x = \Sigma_y = \Sigma = C_0 C_0^\top + \Psi \succ 0$. Thus $T = \Sigma^{-1/2} C_0 C_0^\top \Sigma^{-1/2}$ is symmetric, and for every v, $v^\top T v = \|C_0^\top \Sigma^{-1/2} v\|_2^2 \geq 0$. It is therefore PSD. The common congruence $M \mapsto \Sigma^{1/2} M \Sigma^{1/2}$ preserves rank and is a bijection on the PSD cone, so the whitened equality also holds for the original shared operator class. Under unequal view covariances, this PSD-preserving congruence no longer applies. □

Planar-rotation example. For $T = \rho R_\theta$ with $0 < \rho \leq 1$ in two whitened dimensions, $R_\theta^\top R_\theta = I_2$, so both singular values of T equal ρ. At rank one, Equation (11) gives $\mathfrak{D}_d^\star(1) = \rho$. Moreover, $\mathrm{sym}(R_\theta) = (R_\theta + R_\theta^\top)/2 = \cos\theta\, I_2$. For $0 \leq \theta \leq \pi/2$, the largest positive eigenvalue of $\mathrm{sym}(T)$ is $\rho\cos\theta$, and Equation (12) gives $\mathfrak{D}_s^\star(1) = \rho\cos\theta$. This proves the endpoint and intermediate-angle claims used to motivate the rotation intervention.

For the latent two-view model with $\mathrm{Cov}(\epsilon_q) = \Psi_q$ and $\mathrm{Cov}(\epsilon_d) = \Psi_d$,

$$\Sigma_x = C_q C_q^\top + \Psi_q, \quad \Sigma_y = C_d C_d^\top + \Psi_d, \quad C = C_q C_d^\top,$$

and hence $T = \Sigma_x^{-1/2} C_q C_d^\top \Sigma_y^{-1/2}$. Its left and right singular vectors are the query and document canonical directions. A singular gap controls plug-in subspace stability through Wedin-type perturbation bounds (Cai & Zhang, 2018); covariance conditioning separately affects the whitening error (Gao et al., 2019).

E A Distribution-Free Bound and Its Limitation

Proof of Proposition 1. Let $W = XD^\top$, assume $\|W\|_F \leq R$ almost surely, and constrain $\|M\|_F \leq B$. Conditional on $W_1, \ldots, W_n$, write $Q = n^{-1}\sum_i \epsilon_i W_i$. Von Neumann's inequality gives

$$\begin{aligned}\sup_{\mathrm{rank}(M)\leq r, \|M\|_F \leq B} \langle M, Q\rangle_F &= B\left(\sum_{j\leq r} \sigma_j(Q)^2\right)^{1/2} \\ &\leq B\|Q\|_F.\end{aligned}$$

Jensen's inequality and independence of Rademacher signs imply

$$\mathbb{E}_\epsilon \|Q\|_F \leq \left(\frac{1}{n^2}\sum_i \|W_i\|_F^2\right)^{1/2} \leq \frac{R}{\sqrt{n}}.$$

The shared class is a subset of the dual class, so its Rademacher supremum cannot exceed the dual one. Together these are exactly the inequalities in Proposition 1.

For completeness, let ℓ be an L-Lipschitz bounded margin loss. The constant $\ell(0)$ can be subtracted without changing the deviation of the empirical risk. Symmetrization bounds the expected uniform deviation of the centered loss class by twice its Rademacher complexity; contraction bounds that loss complexity by a constant multiple of L times the linear-score complexity. Applying Proposition 1 gives $O(LBR/\sqrt{n})$ for either family (Bartlett & Mendelson, 2002).

This argument cannot produce the strict penalty in Equation (2): it throws away the leading-r spectrum of Q by upper-bounding it with the full Frobenius norm. More fundamentally, the boundedness assumptions allow a degenerate design $W_i = 0$ (and nonzero designs of arbitrarily small scale), so they cannot imply a uniformly positive complexity gap that depends only on p and r. The local theorem adds the regular design and Gaussian localization needed to expose effective dimension. □

F Experimental Protocols and Supporting Evidence

Retrieval metric. For an evaluated query q, let $z_{qj} \in \{0, 1\}$ indicate relevance at rank j and let $R_q > 0$ be its number of relevant documents. We report the mean over test queries of

$$\mathrm{NDCG@10}(q) = \frac{\sum_{j=1}^{10} z_{qj}/\log_2(j+1)}{\sum_{j=1}^{\min(10,R_q)} 1/\log_2(j+1)}.$$

Synthetic retrieval and local calibration (RQ1). The two-view retrieval simulation uses ambient dimension $p = 32$, latent rank eight, query and document noise standard deviation 0.6, and master seed 20260809. The document loading rotates relative to the query loading through angles $0, 0.2, \ldots, 1.2$ radians. Training sizes are 32–2048 pairs, fitted ranks are $4/8/16$, and five paired seeds control training and test draws. Each positive document is ranked against 299 independent negatives. Figure 3 displays the rank-eight boundary from this grid.

The local-risk calibration uses $p = 12$, $r = 3$, $\sigma = 1$, and $n \in \{64, 256, 1024, 4096\}$. At each of four directional-signal levels, 20,000 nonlinear-projection draws estimate the two risks; 200,000 Gaussian sequence draws estimate SURE selection probability. Table 2 reports the $n = 4096$ comparisons used in RQ1.

Table 2: Local risk and SURE calibration. Entries are simulation/theory; risks are scaled by n.

$\delta^2/(\sigma^2 k)$	Shared risk	Dual risk	$\Pr(\widehat{\pi} = d)$
0.0	32.97/33	63.03/63	0.001/0.001
0.5	48.05/48	63.05/63	0.094/0.093
1.0	62.93/63	62.97/63	0.473/0.474
2.0	93.10/93	63.15/63	0.970/0.969

Training-score diagnostic (RQ2). The rank-eight simulation compares raw plug-in PSD distance with the known population distance and its uncertainty-corrected estimate. On frozen real embeddings, Figure 4B compares raw and corrected training scores in 840 operator fits from 30 available task–encoder pairs. The ten tasks are Climate-FEVER, CQA-English, CQA-TeX, DBPedia Entity, FEVER, HotpotQA, MS MARCO, NQ, TREC-COVID, and Touché-2020. Fits vary rank over $4/8/16/32$, use three seeds and available nested training sizes, and share the four encoders specified in Section 4. Disjoint held-out relevance moments define the Shared-minus-Dual operator-loss advantage against which the training-only scores are compared.

F.1 Real-embedding retrieval and operator-risk protocols

Full-corpus query rotation (RQ3). The reported rotation experiment uses FEVER, HotpotQA, NQ, MS MARCO, and CQA-TeX with BGE-base and GTE-base. Their train/validation/test query counts and corpus sizes are, respectively, 109,810/6,666/6,666 and 5,416,568; 85,000/5,447/7,405 and 5,233,329; 2,071/691/690 and 2,681,468; 502,939/6,980/43 and 8,841,823; and 1,744/581/581 and 68,184. For each dataset–encoder pair, eight relevance-informed planes define query rotations at $0°, 15°, \ldots, 90°$. Documents, relevance labels, and query splits remain fixed across angles.

Shared and Dual rank-16 adapters use the same 1,024 training queries, three paired seeds, epoch permutations, in-batch negatives, and eight frozen hard negatives. AdamW runs for 30 epochs with batch size 256, temperature 0.05, weight decay 10^{-5}, and gradient norm 5. Validation NDCG@10 chooses the learning rate from $\{3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$ and the checkpoint for each family. Test scores use exact float32 full-corpus inner-product search. These are the conditions plotted in Figure 5.

Rank-by-data retrieval (RQ3). The four grids in Figure 6 are MS MARCO with GTE-base, BGE-base, and Contriever-MSMARCO at $90°$, and NQ with GTE-base at $75°$. Each grid crosses seven training sizes $\{32, 64, 128, 256, 512, 1024, 2048\}$ with ranks $\{4, 8, 16, 32\}$ and three paired seeds. Training lasts 30 epochs at learning rate 3×10^{-4}; validation selects the checkpoint and the test ranks the complete corpus.

Held-out operator risk (RQ4). The 168 comparable sample-size trajectories in Figures 8 and 9 come from FEVER, NQ, ArguAna, and SciFact with all four base encoders and three seeds. SciFact (647/162/300 train/validation/test queries, 5,183 documents) and FEVER use their official splits. NQ and ArguAna use deterministic adaptation splits of 2,071/691/690 over 2,681,468 documents and 844/281/281 over 8,674 documents, respectively. One ArguAna test qrel has no released document, leaving 280 evaluable test queries. The available ranks are $4/8/16/32$; ArguAna and SciFact additionally use $64/128/256$.

Encoder-prescribed query and passage prefixes are applied, texts are truncated at 256 tokens, and normalized 768-dimensional embeddings are cached in float16. For each query, the first labeled positive and highest-ranked unlabeled raw candidate form a relevance triple. The held-out triples define a disjoint target moment. Truncated SVD and positive eigentruncation of the training moment give the Dual and Shared operators; squared Frobenius errors are divided by the squared norm of the held-out target. Each trajectory fixes dataset, encoder, rank, and seed and connects only measured nested sample sizes.

Five-fold geometry selection (RQ4). The evaluation behind Table 1 uses ArguAna, MS MARCO, FEVER, NQ, and SciFact with all four base encoders. For each dataset, labeled training-query IDs are sorted, permuted with seed 20260923, and partitioned into five balanced held-out folds shared across encoders. Each fold's target moment uses only held-out queries; the remaining four folds supply training queries. A fold-specific permutation with seed $20260924 + \text{fold}$ provides nested samples of sizes $32, 64, 128, 256, 512, 1024, 2048, 4096$ when available, plus the full remaining pool when it has at most 4096 queries. All runs use ranks $4/8/16/32$, a fixed 256-dimensional selector projection, and 20 internal half-splits. The 2,800 configuration evaluations are nested within 100 encoder–dataset outer folds; regret is averaged across sample-size–rank cells, then held-out folds, as specified in Section 4.

G Additional related work

Hard-negative mining and data augmentation are central to dense retrievers (Xiong et al., 2021; Qu et al., 2021); domain-adaptive and expansion variants include Contriever, GPL, COCO-DR, and HyDE (Izacard et al., 2022; Wang et al., 2022a; Yu et al., 2022; Gao et al., 2023). Query-side anisotropy, compression, and frozen-embedding adapters address different aspects of representation mismatch (Li et al., 2021; Ma et al., 2021; Yoon et al., 2024a;b). CCA and deep CCA learn two-view directions (Hotelling, 1936; Bach & Jordan, 2005; Andrew et al., 2013; Dorfer et al., 2018); perturbation theory studies their sampling stability (Davis & Kahan, 1970; Gao et al., 2019; Cai & Zhang, 2018). Low-rank and nearest-PSD projections (Eckart & Young, 1936; Higham, 1988), fixed-rank manifolds (Vandereycken, 2013; Vandereycken et al., 2013), and risk estimation (Gavish & Donoho, 2017; Mazumder & Weng, 2020) supply ingredients for our operator analysis. Pairwise-ranking theory and NDCG address a different order-based objective (Burges et al., 2005; Clémençon & Vayatis, 2009; Järvelin & Kekäläinen, 2002).

H Identifiability and boundary cases

Marginal mismatch is neither necessary nor sufficient. If $Y = cX$ with $c > 0$ and scoring uses cosine, query and document covariances differ by c^2 while every paired cosine and ranking is unchanged. Conversely, if $X \sim \mathcal{N}(0, I)$ and $Y = RX$ for a nonsymmetric orthogonal rotation R, the marginals agree but the cross-covariance is generally outside the PSD cone. Relevance depends on paired structure.

A bilinear operator does not determine projected dual cosine. For invertible Q, (A, B) and $(QA, Q^{-\top}B)$ preserve $A^\top B$ but change projected norms. Take $M = I$, $x = (1, 1)$, $d_1 = (1, 0)$, and $d_2 = (.9, .9)$. With $A = B = I$, projected cosine ranks d_2 above d_1; with $A = \mathrm{diag}(10, 1)$ and $B = \mathrm{diag}(.1, 1)$, still $A^\top B = I$, but the ranking reverses. The same invariance changes $\|A - B\|_F$ arbitrarily. Operator-level results require bilinear scoring or additional norm control.

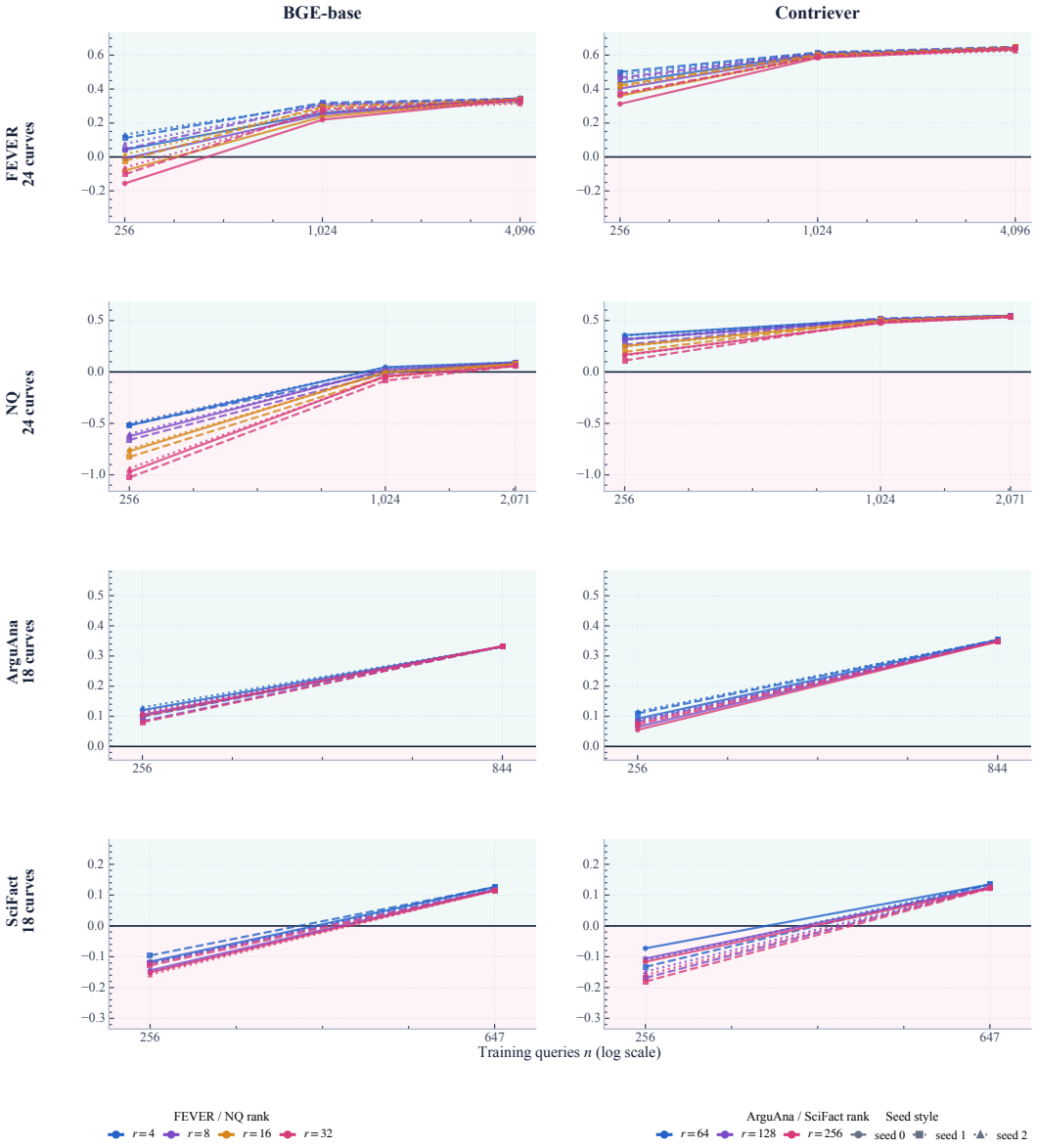

Figure 8: Held-out operator-risk trajectories for BGE-base and Contriever (84 curves). Rows are FEVER and NQ (24 curves each), followed by ArguAna and SciFact (18 each). Each line fixes a dataset, encoder, rank, and seed and connects measured nested training sizes. Color denotes rank; line style denotes seed. Positive normalized Shared-minus-Dual loss favors Dual.

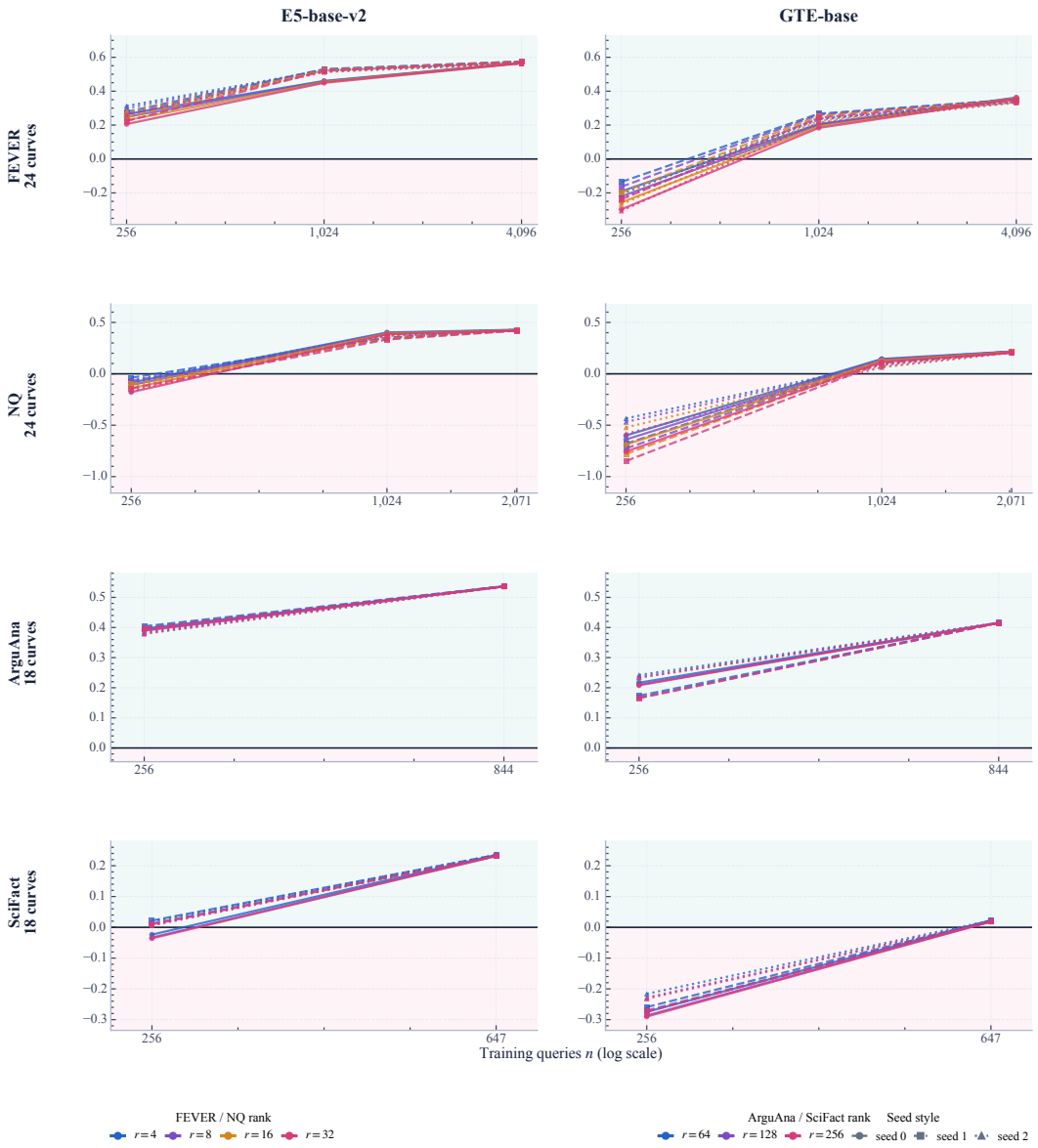

Figure 9: The remaining 84 held-out operator-risk trajectories, for E5-base-v2 and GTE-base. Grouping, axes, colors, and line styles match Figure 8.