RELEVANCE IS NOT VARIANCE: LOCAL FISHER-SPECTRAL ADAPTATION FOR DENSE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) relies on relevant external evidence to ground language model responses. Retrieving this evidence remains challenging when a fixed embedding space must serve queries with different information needs. Directions that capture substantial variation in that space may offer little help in distinguishing relevant evidence from distractors. We study how relevance feedback can guide query adaptation while keeping the encoder and document index unchanged. We propose Fisher-Spectral Adaptation (FISA), a supervised adapter that uses the initial retrieval results to identify locally discriminative directions and refine the query for a second search. Experiments with five frozen embedding models on four BEIR tasks show a 3.63-point gain in macro nDCG@10 (8.58% relative) over the strongest deterministic baseline, with improvements in 16 of 20 encoder–dataset comparisons. The gains are stable across 20 training seeds, while evaluation on an unseen domain reveals limits to transfer. These results demonstrate the potential of adapting retrieval geometry to query-specific evidence.

1 INTRODUCTION

Retrieval-augmented generation (RAG) augments a language model with evidence retrieved from an external corpus, allowing the system to answer questions that are private, recent, or too detailed to be stored reliably in model parameters (Lewis et al., 2020). Although a RAG pipeline is often evaluated by its final answer, retrieval determines the evidence that the reader can see in the first place. When the retriever misses a relevant passage or ranks distracting passages ahead of it, a stronger generator cannot fully repair the information bottleneck. Improving retrieval is therefore a direct way to strengthen RAG without retraining the reader or expanding its context window.

Dense retrieval provides an appealing systems interface for this purpose. A query is mapped to one vector, documents are stored once in a shared index, and maximum-inner-product search returns evidence at scale (Reimers & Gurevych, 2019; Karpukhin et al., 2020; Xiong et al., 2021; Izacard et al., 2022). This modularity also creates a restrictive geometric assumption: the same frozen embedding space must serve every query. The axes that encode useful distinctions for a scientific claim may not be the axes that separate financial evidence or opposing arguments. A single global geometry can therefore be well trained on average while remaining locally ambiguous for a particular information need.

The distinction between variance and relevance is especially important within a retrieved neighborhood. Highly ranked documents can share vocabulary, topic, or writing style even when they differ in whether they answer the query. A direction that explains much of their overall variation may therefore separate document types without separating useful evidence from distractors. Conversely, a direction with modest variance can be informative when relevant and irrelevant candidates differ consistently along it. This motivates evaluating local directions through both between-group separation and within-group dispersion, while learning how much trust to place in the feedback itself.

Existing post-hoc approaches attack different parts of this problem. Corpus-level spectral methods reshape one geometry for all queries (Mu & Viswanath, 2018; Li et al., 2020; 2026), while query-aware dimension methods adapt weights within a fixed coordinate system (D'Erasmo et al., 2025; Wu et al., 2026; Yang & Zhao, 2026). Dense pseudo-relevance feedback instead uses first-pass candidates to modify the query (Rocchio, 1971; Yu et al., 2021; Wang et al., 2023b), but typically does not ask which local directions actually separate useful evidence from hard negatives. These approaches

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.

therefore adapt the weighting, the feedback signal, or the global geometry, but not the coordinate system in which relevance is expressed for the current neighborhood. The central question of this work is thus not only *which coordinates should matter*, but also *which coordinate system best exposes relevance for this query*.

We propose **Fisher-Spectral Adaptation (FISA)**, a supervised query adapter that uses the first retrieval pass to identify which directions matter for the current information need. It combines query-conditioned coordinate reweighting with a local Fisher criterion: directions are useful when they separate learned positive and negative evidence relative to the variation within each group. A confidence gate controls the resulting query update before a second search over the full corpus. The document embeddings and index remain unchanged, and the second search can retrieve documents outside the initial neighborhood.

Our contributions are threefold:

- A query-local view of retrieval geometry treats the coordinate system for interpreting feedback as an adaptive component. This perspective evaluates directions by their discriminative value for the current information need, guiding geometric adaptation of frozen retrievers.
- **FISA** realizes this view through a Fisher-style correction that combines a local spectral basis with learned positive and negative evidence. Directions are evaluated by between-group separation relative to within-group variation, linking candidate geometry to predicted relevance.
- A bounded adapter combines query-conditioned coordinate reweighting with confidence-controlled residual updates to regulate the initial query and the acceptance of local feedback. The resulting query supports a second full-corpus search with the encoder and document index fixed.

2 Related Work

2.1 Post-hoc geometry for dense representations

One line of work asks *where* relevance should be represented in a frozen embedding space. Anisotropy and local manifold structure motivate global post-processing (Ethayarajh, 2019; Cai et al., 2021): dominant-component removal, BERT-flow, whitening, and SpecTemp reshape one geometry for all queries (Mu & Viswanath, 2018; Li et al., 2020; Su et al., 2021; Li et al., 2026). Learned alternatives construct reusable transformations or nested prefixes (Kusupati et al., 2022; Yoon et al., 2024a;b). DIME and its subsequent reproducibility, pruning, and risk analyses instead estimate query-specific importance in a fixed coordinate system (Faggioli et al., 2024; Campagnano et al., 2025; Faggioli et al., 2026; D'Erasmo et al., 2026); ECLIPSE-PRF, Learning to Select, and AdaWidth further introduce contrastive, supervised, or adaptive-width decisions (D'Erasmo et al., 2025; Wu et al., 2026; Yang & Zhao, 2026). These methods make a transform, weight, or width adaptive, whereas FISA reconstructs the basis itself from the current query neighborhood.

2.2 Relevance feedback and two-pass dense retrieval

The complementary line asks *where query-specific evidence comes from*. Rocchio, neural PRF, and contextual query expansion revise a query or ranking using early results (Rocchio, 1971; Li et al., 2018; Zheng et al., 2020). ANCE-PRF keeps the document index fixed, TOUR optimizes one query vector from test-time pseudo labels, and broader studies document both dense-PRF gains and query drift (Yu et al., 2021; Sung et al., 2023; Li et al., 2023). ColBERT instead uses multiple token vectors, with ColBERT-PRF and contrastive weighting selecting or expanding feedback embeddings (Khattab & Zaharia, 2020; Wang et al., 2023b;a). FISA retains the single-vector interface: the first pass supplies learned positive and negative evidence for a local basis, and the resulting query then searches the full unchanged corpus. Feedback therefore provides the evidence, while local geometry determines how that evidence changes the query.

Because FISA learns from source-domain relevance labels, comparisons with unsupervised feedback methods evaluate the complete trained system. We therefore use ablations with matched supervision to assess the contribution of local geometry and learned evidence. At inference, the adapter observes only frozen embeddings and retrieval scores; it receives no relevance labels and must estimate whether the first-pass neighborhood supports a useful update.

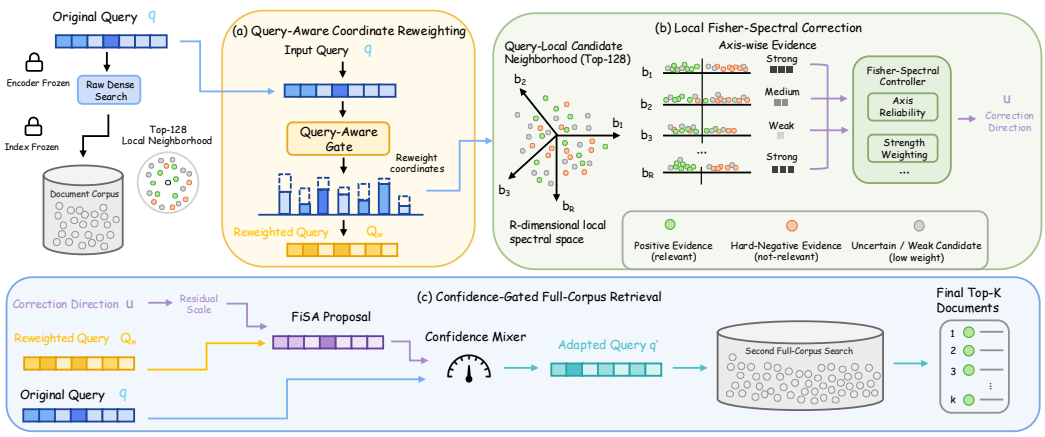


Figure 1: Overall framework of FISA.

3 METHODOLOGY

The design follows three constraints. First, a query-only path should provide a safe correction before noisy feedback is interpreted. Second, first-pass candidates should identify a relevance-separating direction rather than merely a high-variance direction. Third, the model should be able to reject that direction when the neighborhood is unreliable. Figure 1 realizes these constraints as coordinate reweighting, a local Fisher-spectral update, and confidence-controlled acceptance.

Formally, let $q \in \mathbb{R}^D$ be a frozen, L2-normalized query embedding and $\mathcal{D} = \{d_i\}_{i=1}^N$ a frozen document index. Raw Dense first retrieves the ordered neighborhood $C_q = \{d_{q1}, \ldots, d_{qL}\}$ with $L = 128$ and scores $s_{qi} = q^\top d_{qi}$. The adapter constructs $q' = A_\theta(q, C_q, s_q)$ and searches the complete corpus again. The first-pass neighborhood is therefore adaptation evidence, not a fixed reranking pool or a recall ceiling; document vectors never change.

The adapter never replaces dense scoring with an unrestricted query–document network. It uses the neighborhood only to construct one vector in the original encoder space, after which ordinary maximum-inner-product search is reused. This separation prevents the controller from becoming a hidden reranker and keeps corpus-dependent computation confined to the two retrieval passes.

3.1 QUERY-CONDITIONED COORDINATE REWEIGHTING

The first stage asks which coordinates of the frozen encoder should influence the current query. It predicts a bounded multiplier for every coordinate rather than dropping dimensions or reducing the embedding width, and a separate adapter is trained for each frozen encoder.

Query-conditioned coordinate context. A lightweight context network ϕ_θ maps the normalized query to $h_q \in \mathbb{R}^H$ ($H = 256$ in our experiments). From this shared context, one head predicts coordinate logits $z_q \in \mathbb{R}^D$ and another predicts a positive, query-dependent temperature τ_q. The context h_q is a low-capacity input to the controller heads, not a compressed replacement for the retrieval representation. The original D-dimensional q remains on the direct path through the coordinate product, the local alignment and projection, and the final residual interpolation. Thus the controllers share a regularized query summary without forcing retrieval through a 256-dimensional bottleneck.

Centering the logits removes a common shift that would move every gate together:

$$g_q = 2\sigma\left(\frac{z_q - \mathrm{mean}(z_q)}{\tau_q}\right), \qquad q_{\mathrm{sel}} = \mathrm{Norm}(q \odot g_q). \tag{1}$$

Because $g_{qj} \in (0, 2)$, a coordinate can be softly suppressed or amplified without an unbounded sign flip. Zero-initialized logits make the stage begin at the identity mapping. A small τ_q creates a sharper coordinate profile, whereas a large value keeps the gates closer to one.

Supervised importance distillation. Training labels provide a teacher signal that is unavailable for a new query. Let μ_q^+ be the graded-relevance-weighted centroid of relevant candidates, and let μ_q^- be a hard-negative centroid whose weights are proportional to first-pass scores among non-relevant candidates. Their query-compatible coordinate evidence is

$$t_q = q \odot (\mu_q^+ - \mu_q^-), \qquad p_q^{\dim} = \mathrm{softmax}(t_q / T_{\dim}). \tag{2}$$

A coordinate receives a strong target only when the query activates it and the relevant and non-relevant centroids differ along it in a compatible direction. The logits z_q are matched to $p_q^{\dim}$ by KL divergence. At inference, the labels and teacher centroids disappear; the network predicts the gate from q alone.

Why a coordinate gate precedes local geometry. The frozen encoder basis remains useful even though it is not sufficient. It contains the representation that produced the first ranking and provides a stable query-only correction when candidate feedback is weak. Applying the gate before constructing local evidence emphasizes query features that should guide candidate interpretation. The next stage can then add directions that need not align with any single encoder coordinate.

3.2 Local Fisher-Spectral Update

The second stage uses the first-pass neighborhood to identify directions that separate useful evidence from distracting evidence. It does not whiten the corpus or compress documents. Instead, it constructs a temporary, query-conditioned coordinate system, estimates a discriminative direction in that system, and maps the direction back to the original embedding space.

Candidate-centered neighborhood. Let $\bar{d}_q = L^{-1} \sum_{i=1}^{L} d_{qi}$ be the candidate centroid and let $X_q \in \mathbb{R}^{L \times D}$ stack the centered candidates $x_{qi} = d_{qi} - \bar{d}_q$. Centering makes the local axes describe how the retrieved documents vary around their own neighborhood rather than around the origin of the global encoder. The query determines which documents enter X_q, while those documents jointly determine the temporary geometry.

Local spectral axes. Rather than decomposing a $D \times D$ covariance matrix, FISA works in the smaller sample space:

$$G_q = \frac{X_q X_q^\top}{L-1} = V_q \Lambda_q V_q^\top, \qquad U_q = X_q^\top V_q \big((L-1)\Lambda_q\big)^{-1/2}. \tag{3}$$

The columns of $U_q \in \mathbb{R}^{D \times R_q}$ are orthonormal axes in the original embedding space, and candidate i has local coordinates $y_{qi} = U_q^\top x_{qi}$. Since the candidates are centered, $R_q \le L - 1 = 127$ even when D is much larger. Numerically unsupported axes are masked, and each eigenvector sign is deterministically anchored so equivalent decompositions do not present arbitrary sign flips to the controller.

The eigenvalues indicate which directions are available in the neighborhood, but they do not establish relevance. A high-variance axis may encode topic, style, or length shared by both useful and distracting documents. FISA therefore treats the local spectrum as context for a discriminative statistic, not as a relevance ranking by itself.

Learned positive and negative evidence. A fixed pseudo-relevant prefix assumes that rank alone identifies supporting documents. FISA instead predicts two independent attention distributions over all L candidates. For candidate i, the attention network combines h_q with the base score s_{qi}, normalized rank, alignment between x_{qi} and $q_{\mathrm{sel}} - \bar{d}_q$, and centered candidate norm. Two logits yield

$$a_{qi}^+ = \mathrm{softmax}_i(\ell_{qi}^+), \qquad a_{qi}^- = \mathrm{softmax}_i(\ell_{qi}^-). \tag{4}$$

These are separate distributions rather than complementary class probabilities: a document may receive little mass from both heads, and the heads may focus on different subsets of hard evidence. Graded gains supervise a^+ during training, while high-scoring non-relevant candidates supervise a^-. Neither head sees qrels at test time.

Fisher-normalized separation. Along each local axis r, the two attention distributions define positive and negative means and variances:

$$\mu_{qr}^{\pm} = \sum_i a_{qi}^{\pm} y_{qir}, \qquad v_{qr}^{\pm} = \sum_i a_{qi}^{\pm} (y_{qir} - \mu_{qr}^{\pm})^2. \tag{5}$$

The signed diagonal Fisher-style statistic is then

$$f_{qr} = \frac{\mu_{qr}^{+} - \mu_{qr}^{-}}{v_{qr}^{+} + v_{qr}^{-} + \lambda}. \tag{6}$$

This is a query-local, soft-evidence analogue of Fisher discrimination (Fisher, 1936; Sugiyama, 2007). The numerator measures positive–negative separation, the denominator discounts separation that is diffuse within either group, and the sign points toward positive evidence. We use the diagonal form because 128 candidates are insufficient for a reliable full within-class inverse. The scalar λ is only a fixed numerical ridge that prevents division by a near-zero variance; it is not learned evidence and does not compete with f.

Spectral control and direction synthesis. The statistic f reports observed separation, but by itself cannot tell whether a low-variance ratio is reliable or compensate for a systematic bias shared by similar training queries. A shared per-axis controller therefore reads h_q, the standardized log eigenvalue, the selected-query coordinate, both evidence means, and their summed variance. It predicts logits $\widetilde{m}_{qr}$ and $\widetilde{\delta}_{qr}$, which are bounded as $m_{qr} = \sigma(\widetilde{m}_{qr})$ and $\delta_{qr} = \tanh(\widetilde{\delta}_{qr})$. The resulting direction is

$$\alpha_{qr} = m_{qr}\big(f_{qr} + \delta_{qr}\big), \qquad u_q = \mathrm{Norm}\left(\sum_{r=1}^{R_q} \alpha_{qr} u_{qr}\right), \tag{7}$$

where u_{qr} is column r of U_q. Here m and δ are per-query, per-axis network outputs rather than manual hyperparameters: m suppresses an unreliable axis and δ supplies a bounded correction, while f keeps the update anchored in observed positive–negative separation. This decomposition is more inspectable than predicting α directly. Mapping through U_q returns $u_q \in \mathbb{R}^D$, so the local coordinates never replace the frozen index.

3.3 Confidence-Gated Residual Retrieval

The final stage distinguishes how far to move along the local direction from how much to trust the complete proposal. The distinction matters because u_q may be informative even when the first-pass neighborhood is not reliable enough to justify replacing the raw query.

Residual proposal. The entropies of the positive attention, negative attention, and coordinate importance distributions summarize how concentrated the intermediate decisions are. Together with h_q, they are passed to a residual head that predicts a bounded scale $\eta_q \in (-1, 1)$:

$$q_{\mathrm{prop}} = \mathrm{Norm}(q_{\mathrm{sel}} + \eta_q u_q). \tag{8}$$

The sign of η_q chooses the orientation of u_q, and its magnitude sets the step along that direction. When $\eta_q \approx 0$, the proposal retains only the coordinate-reweighting stage; the bound prevents an arbitrarily large pre-normalization residual.

Confidence-controlled acceptance. A separate head receives h_q together with the first-pass top score, top-one/top-two margin, and score dispersion. It predicts $c_q = \sigma(\rho_q) \in (0, 1)$ and interpolates between the raw query and the full proposal:

$$q' = \mathrm{Norm}\left(q + c_q(q_{\mathrm{prop}} - q)\right). \tag{9}$$

Small c_q keeps retrieval close to the direct D-dimensional raw-query path; large c_q accepts both coordinate and spectral changes. The two scalars are therefore not redundant: η_q controls motion inside the proposed update, whereas c_q decides whether that proposal as a whole should override q. Since a sigmoid never reaches zero at finite parameters, Raw Dense is a limiting fallback rather than a discrete switch. The confidence bias is initialized to -2, so optimization begins near the raw-query path and must learn to accept larger changes from evidence.

Second full-corpus retrieval. The adapted query is scored against the original frozen index, $\mathrm{score}(q', d_i) = q'^{\top} d_i$, and the system returns the final top-1000. This is not fixed-candidate reranking: the top-128 documents are used to infer q', but any corpus document can enter the second ranking. The method therefore preserves the single-vector dense-retrieval interface while allowing the first pass to retrieve evidence that guides a new full-corpus search.

Learning the adapter. The main objective is listwise cross-entropy over graded candidate gains. Auxiliary terms distill the coordinate target in Equation 2, supervise both evidence distributions, match the local direction to a label-derived Fisher target, regularize the residual, and preserve cosine similarity with the raw query:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rank}} + \mathcal{L}_{\mathrm{dim}} + 0.1\mathcal{L}_{\mathrm{dir}} + 0.05\mathcal{L}_{\mathrm{attn}} + 0.2\mathcal{L}_{\mathrm{res}} + 0.2\mathcal{L}_{\mathrm{pres}}. \tag{10}$$

Labels supervise only training. Training queries are included only when their first-pass top-128 contains at least one judged relevant document, because the supervised positive target is otherwise undefined. Development evaluation, by contrast, retains every query and performs the complete second search. We train one adapter per frozen encoder and select checkpoints by full-corpus development nDCG@10, preferring epochs that do not regress on any of the three source development domains. Exact network inputs, teacher targets, initialization, checkpoint fallback, and numerical safeguards appear in Appendix A.

Separation of responsibilities. The stages are deliberately not collapsed into one black-box scorer. Coordinate reweighting provides a query-only correction learned from source-domain regularities. The Fisher-spectral stage uses candidates only to estimate a direction, leaving the final scoring rule linear and the document index reusable. The confidence stage controls whether candidate evidence is reliable enough to override the raw query. This decomposition makes failures observable: the model can fail because encoder coordinates are poorly weighted, because the first-pass neighborhood suggests the wrong local direction, or because a useful proposal is accepted too weakly or too aggressively. The paired ablations in Section 4 remove these capabilities one at a time while keeping the remaining training protocol fixed.

Computation and invariants. For fixed L, local basis construction costs $O(L^2D + L^3)$ and stores $O(LD + L^2)$ temporary values. With $L = 128$, this adaptation cost does not grow with corpus size; the corpus-dependent addition is the second maximum-inner- product search. Candidate centering, local coordinates, and attention statistics are ephemeral: they never replace or re-index document embeddings. At inference, the adapter receives only the raw query, first-pass scores and ranks, and the retrieved candidate embeddings; no relevance label, teacher centroid, or corpus-fitted local basis is available. Its output is one normalized vector in the original encoder space, so the document embeddings, ANN index, and downstream RAG interface remain unchanged.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Benchmarks. We evaluate on four BEIR retrieval tasks (Thakur et al., 2021): NFCORPUS, SCIFACT, FIQA, and ARGUANA. NFCORPUS contains biomedical information needs with graded relevance; SCIFACT retrieves scientific evidence for claims; FIQA covers financial questions; and ARGUANA matches arguments with counterarguments. The first three datasets supply official training queries. We create deterministic 90/10 train–development partitions inside those training splits and evaluate on their official test queries. ARGUANA contributes no training queries, development labels, fabricated split, or basis-fitting documents; it is used only as an unseen-domain test.

Encoders and methods. The frozen embedding models are BGE-base, E5-base, BGE-large, Qwen3-Embedding-0.6B, and BGE-M3 (Wang et al., 2022; Xiao et al., 2023; Chen et al., 2024; Zhang et al., 2025). Model revisions, pooling rules, and query/document prefixes are pinned. We compare Raw Dense, SPECTEMP-256 (Li et al., 2026), ECLIPSE-PRF (D'Erasmo et al., 2025), and FISA. Raw Dense, SPECTEMP-256, and ECLIPSE-PRF are evaluated once per encoder–dataset cell, whereas FISA uses 20 predeclared seeds per encoder, yielding 400 physical test results.

Table 1: nDCG@10 (%) across five frozen encoders and four BEIR tasks. FISA is averaged over 20 seeds; best and second-best results within each encoder–dataset comparison are bold and underlined.

Method	Encoder	NFCORPUS	SCIFACT	FIQA	ARGUANA	Avg.
Raw Dense	BGE-base	37.56	<u>74.04</u>	<u>40.62</u>	19.31	42.88
	E5-base	35.60	<u>71.94</u>	<u>39.87</u>	12.01	39.86
	BGE-large	38.38	74.63	45.00	20.02	44.51
	Qwen3-0.6B	36.45	<u>70.23</u>	<u>47.00</u>	21.16	43.71
	BGE-M3	31.60	<u>64.37</u>	<u>41.30</u>	15.47	38.18
SPECTEMP-256	BGE-base	36.32	72.80	37.16	18.70	41.25
	E5-base	32.40	69.66	37.38	12.04	37.87
	BGE-large	36.82	72.90	42.92	18.98	42.91
	Qwen3-0.6B	35.29	70.07	46.02	20.42	42.95
	BGE-M3	30.22	63.30	39.88	15.50	37.22
ECLIPSE-PRF	BGE-base	<u>38.63</u>	73.72	40.18	**20.17**	<u>43.18</u>
	E5-base	<u>36.77</u>	71.07	39.38	**16.97**	<u>41.05</u>
	BGE-large	<u>39.22</u>	<u>74.74</u>	<u>45.51</u>	**21.14**	<u>45.15</u>
	Qwen3-0.6B	<u>38.42</u>	69.64	46.30	**22.10**	<u>44.12</u>
	BGE-M3	<u>32.24</u>	63.01	41.24	<u>16.32</u>	<u>38.20</u>
FISA	BGE-base	**39.79**	**84.65**	**41.27**	<u>19.74</u>	**46.36**
	E5-base	**38.90**	**83.08**	**42.16**	<u>15.39</u>	**44.88**
	BGE-large	**40.88**	**83.48**	**46.09**	<u>20.53</u>	**47.74**
	Qwen3-0.6B	**38.87**	**82.28**	**47.27**	<u>21.97</u>	**47.60**
	BGE-M3	**34.73**	**78.89**	**42.89**	**16.57**	**43.27**

Metrics and inference. The primary metric is nDCG@10; we additionally report MRR@10, Recall@100, and Recall@1000. Main paired comparisons use 10,000-draw hierarchical bootstrap (Efron & Tibshirani, 1994): query clusters are shared across methods and encoders, while learned-model seed clusters are resampled separately for each encoder. Paired query-cluster randomization tests assess nDCG@10 with Holm correction. The five component ablations use the same seeds $\{13, 21, 42, 87, 100\}$ for Full FISA and every variant.

4.2 MAIN RESULTS

Aggregate effectiveness. Averaged equally over the 20 encoder–dataset cells, FISA reaches 0.4597 nDCG@10, compared with 0.4234 for ECLIPSE-PRF, 0.4183 for Raw Dense, and 0.4044 for SPECTEMP-256. Table 2 also shows the strongest macro MRR@10, Recall@100, and Recall@1000. Against ECLIPSE-PRF, the gains in nDCG@10 and MRR@10 are 3.63 and 4.25 points, whereas Recall@100 and Recall@1000 improve by 0.32 and 0.26 points. The Recall@100 confidence interval crosses zero; the Recall@1000 interval remains positive but small. The evidence therefore points primarily to better ranking, with more limited gains in the coverage of already retrievable documents.

Table 2: Macro effectiveness across five encoders and four datasets. FISA reports the 20-seed mean and sample standard deviation; metrics are on a 0–1 scale.

Method	nDCG@10	MRR@10	Recall@100	Recall@1000
Raw Dense	.4183	.4557	.7438	.8878
SPECTEMP-256	.4044	.4383	.7348	.8789
ECLIPSE-PRF	.4234	.4495	.7586	.8939
FISA	**.4597±.0006**	**.4920±.0011**	**.7618±.0014**	**.8964±.0009**

Paired comparisons. The primary-metric improvement over ECLIPSE-PRF is 8.58% relative, with a hierarchical 95% confidence interval of $[0.02942, 0.04331]$. Table 3 reports positive intervals against all three baselines and Holm-adjusted randomization p-values of 0.00030. These comparisons account for shared test queries and training-seed variation; they are conditional on the evaluated encoders and tasks.

Variation across encoders and domains. Table 1 shows improvements over Raw Dense and SPECTEMP-256 in all 20 cells and over ECLIPSE-PRF in 16. The largest gains occur on SCIFACT, where the mean margin over ECLIPSE-PRF is 12.04 nDCG@10 points and every encoder benefits.

Table 3: Paired nDCG@10 comparisons of FISA against the three deterministic baselines. Non-wins count encoder–dataset cells where FISA does not improve on the baseline.

Baseline	Δ	Hierarchical 95% CI	Holm p	Non-wins
Raw Dense	.04143	[.03473,.04833]	.00030	0/20
SPECTEMP-256	.05532	[.04848,.06239]	.00030	0/20
ECLIPSE-PRF	.03633	[.02942,.04331]	.00030	4/20

NFCORPUS and FIQA show smaller but consistently positive margins. On unseen ARGUANA, FISA remains above Raw Dense and SPECTEMP-256 for all five encoders, but ECLIPSE-PRF is stronger on four; only BGE-M3 favors FISA. The mean deficit is 0.50 nDCG@10 points. This limits the transfer claim: a local geometry learned on the source domains can be too specific for argument retrieval, where simpler unsupervised feedback can transfer better. Appendix E retains the dataset-level secondary metrics, encoder averages, and cell-level seed dispersion.

4.3 ABLATION STUDY

Table 4: Five-seed component ablation in nDCG@10 (%). Δ is the macro change from Full FISA; all variants share encoders, splits, seeds, and checkpoint selection.

Variant	NFCORPUS	SCIFACT	FIQA	ARGUANA	Avg.	Δ
Full FISA	**38.57**	**82.37**	**43.95**	18.85	**45.93**	—
w/o Coordinate Reweighting	37.29	71.69	43.60	19.19	42.94	-2.99
Fixed PRF Attention	37.29	80.74	43.53	17.41	44.74	-1.19
Dimension Only	36.87	81.60	43.24	17.44	44.79	-1.15
Global Basis	38.36	81.97	43.39	**19.42**	45.79	-0.15
w/o Trust Region	37.37	80.13	42.76	18.30	44.64	-1.29

Coordinate reweighting is the largest contributor in Table 4: removing it costs 2.99 points, largely because the adapter loses its strong SCIFACT behavior. Dimension Only already outperforms the deterministic baselines in the paired five-seed macro, but Full FISA adds another 1.15 points, showing that candidate-conditioned Fisher-spectral correction contributes beyond query-only reweighting. Replacing learned evidence with fixed score-based PRF attention loses 1.19 points and Full FISA wins all 20 paired encoder–dataset cells, so the local basis does not remove the need to distinguish supporting evidence from hard negatives.

The remaining variants qualify the geometric claim. Replacing the query-local basis with a source-corpus global rank-128 PCA basis loses only 0.15 points on average, its hierarchical interval includes zero, and it is stronger on ARGUANA. Locality is therefore helpful in the complete system but not a universal theorem about every domain. By contrast, removing confidence-controlled acceptance loses 1.29 points and increases seed variability, supporting the role of conservative fallback when candidate evidence is noisy. Full definitions, intervals, adjusted p-values, and non-win counts are reported in Appendix F.

4.4 EFFICIENCY ANALYSIS

FISA leaves document embeddings and index storage unchanged. Its online addition is a decomposition of a 128×128 candidate Gram matrix, small MLP controllers, and one additional full-corpus search. For fixed candidate count, local adaptation costs $O(L^2D + L^3)$ and does not scale with corpus size; only the second maximum-inner-product search does. This makes the method attractive when a large shared index is expensive to rebuild, but less attractive under an absolute single-pass latency budget.

The saved experiment logs do not impose one comparable end-to-end timing boundary across methods: encoding, index construction, cached first-pass mining, adaptation, transformed-index creation, and second-pass search are included differently. We therefore do not claim latency or throughput superiority. The supported systems claim is narrower: FISA improves retrieval while retaining the frozen index and the same downstream reader interface.

4.5 TRANSFER, STABILITY, AND COMPARISON SCOPE

Seen and unseen behavior. The three source-domain test sets and the unseen ARGUANA result should not be read as one homogeneous average. On the source domains, Full FISA consistently combines the query gate with candidate-conditioned correction. On ARGUANA, both ECLIPSE-PRF and the Global Basis ablation become more competitive, suggesting that the local candidate geometry learned from biomedical, scientific, and financial supervision transfers imperfectly to argument–counterargument matching. Importantly, the confidence gate does not fully solve this shift: it reduces harmful drift on average, but it is learned from source-domain score patterns. A stronger deployment rule would explicitly compare raw, global-basis, and local proposals before selecting one.

Seed stability. The 20-seed result is not driven by the original five runs. Their macro nDCG@10 is 0.45935, while the later 15-seed extension gives 0.45983; the difference is 0.00049. Cell-level standard deviations are larger than the aggregate deviation because averaging over encoders and datasets cancels some variation, but the largest cell standard deviation is still below 0.0055. The main conclusion is therefore stable at both the macro and cell levels. We nevertheless keep the baseline variance at zero rather than duplicating deterministic outputs under seed labels, because such duplication would manufacture uncertainty that was not physically observed.

What the external comparison establishes. FISA uses source-domain relevance supervision, whereas Raw Dense, SPECTEMP-256, and the inference rule of ECLIPSE-PRF do not use equivalent training labels. The main table therefore evaluates the complete trained adapter; it does not isolate supervision from architecture. The paired ablations provide the controlled evidence for the three-stage design because every variant shares training queries, seeds, optimizer, checkpoint rule, and test protocol. In particular, the gain over Dimension Only attributes additional value to candidate feedback under matched supervision, while Fixed PRF Attention isolates the benefit of learning evidence assignments rather than relying on score order.

Retrieval rather than generation. The study measures which evidence is retrieved, not how a language model uses it. Higher nDCG@10 and recall can make better evidence available to a downstream RAG reader, but they do not guarantee factual generation, correct citation, or robust multi-document synthesis. We keep this boundary explicit because the frozen-index interface makes a clean downstream study possible: the same reader can be run with Raw Dense and FISA contexts without retraining either the reader or the document encoder.

5 CONCLUSION

We introduced FISA, a three-stage adapter for two-pass dense retrieval. The method first reweights the coordinates of a frozen query representation, then uses the first-pass candidates to construct local spectral axes and measure positive–negative separation relative to within-group variation. A confidence gate controls how much of the resulting proposal is accepted before a second full-corpus search. This design keeps the document index fixed while allowing relevance geometry to depend on the current query.

Across five encoders and four tasks, FISA provides a stable overall gain and wins 16 of 20 comparisons against the strongest deterministic baseline. The ablations show that coordinate reweighting supplies the largest single component, while learned evidence, Fisher-spectral correction, and conservative acceptance all add value. The ARGUANA exception marks the transfer boundary: a source-trained local basis is not always safer than a simpler global alternative. This motivates a deployment rule that chooses among raw, global, and local proposals; whether the retrieval gains improve grounded generation remains a separate downstream question.

AI USE STATEMENT

Generative AI tools were used for code generation and refactoring, debugging assistance, literature search, and language editing. The authors verified all code, citations, results, and claims and take full responsibility for the submission.

REPRODUCIBILITY STATEMENT

The supplementary artifact contains the shared scientific implementation, formal configurations, exact model revisions, dataset identifiers, 20 main seeds and five paired ablation seeds, preserved implementation sources and checkpoint identifiers, and scripts that reproduce the main and ablation tables and paired statistical tests from saved per-query records. Appendices A–G record the method parameterization, worked example, experimental protocol, complete results, ablation definitions, and integrity checks.

REFERENCES

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xYGNO86OWDH.

Cesare Campagnano, Antonio Mallia, and Fabrizio Silvestri. Unveiling DIME: Reproducibility, generalizability, and formal analysis of dimension importance estimation for dense retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3367–3376, 2025. doi: 10.1145/3726302.3730318.

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. URL https://arxiv.org/abs/2402.03216.

Giulio D'Erasmo, Giovanni Trappolini, Fabrizio Silvestri, and Nicola Tonellotto. ECLIPSE: Contrastive dimension importance estimation with pseudo-irrelevance feedback for dense retrieval. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval*, pp. 147–154, 2025. doi: 10.1145/3731120.3744579.

Giulio D'Erasmo, Cesare Campagnano, Antonio Mallia, Pierpaolo Brutti, Nicola Tonellotto, and Fabrizio Silvestri. Statistical foundations of DIME: Risk estimation for practical index selection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 722–730, 2026. doi: 10.18653/v1/2026.eacl-short.51.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 55–65, 2019. doi: 10.18653/v1/D19-1006.

Guglielmo Faggioli, Nicola Ferro, Raffaele Perego, and Nicola Tonellotto. Dimension importance estimation for dense information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1318–1328, 2024. doi: 10.1145/3626772.3657691.

Guglielmo Faggioli, Nicola Ferro, Raffaele Perego, and Nicola Tonellotto. Getting off the DIME: Dimension pruning via dimension importance estimation for dense information retrieval. *ACM Transactions on Information Systems*, 44(1):9:1–9:34, 2026. doi: 10.1145/3765619.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

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.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, 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.

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

Canjia Li, Yingfei Sun, Ben He, Le Wang, Kai Hui, Andrew Yates, Le Sun, and Jungang Xu. NPRF: A neural pseudo relevance feedback framework for ad-hoc information retrieval. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4482–4491, 2018. doi: 10.18653/v1/D18-1478.

Hang Li, Ahmed Mourad, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls. *ACM Transactions on Information Systems*, 41(3):62:1–62:40, 2023. doi: 10.1145/3570724.

Yongkang Li, Panagiotis Eustratiadis, and Evangelos Kanoulas. Spectral tempering for embedding compression in dense passage retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3941–3946, 2026. doi: 10.1145/3805712.3809896.

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

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Joseph John Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313–323. Prentice Hall, 1971.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*, 2021. URL https://arxiv.org/abs/2103.15316.

Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, 2007. URL https://www.jmlr.org/papers/v8/sugiyama07b.html.

Mujeen Sung, Jungsoo Park, Jaewoo Kang, Danqi Chen, and Jinhyuk Lee. Optimizing test-time query representations for dense retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5731–5746, 2023. doi: 10.18653/v1/2023.findings-acl.354.

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.

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. URL https://arxiv.org/abs/2212.03533.

Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. Effective contrastive weighting for dense query expansion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12688–12704, 2023a. doi: 10.18653/v1/2023.acl-long.710.

Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. ColBERT-PRF: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Transactions on the Web*, 17 (1):3:1–3:39, 2023b. doi: 10.1145/3572405.

Zhanyu Wu, Richong Zhang, and Zhijie Nie. Learning to select: Query-aware adaptive dimension selection for dense retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18666–18677, 2026. doi: 10.18653/v1/2026.acl-long.849. URL https://aclanthology.org/2026.acl-long.849/.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*, 2023. URL https://arxiv.org/abs/2309.07597.

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

Shubing Yang and Dongfang Zhao. AdaWidth: Query-adaptive embedding width for dense retrieval. *arXiv preprint arXiv:2608.23862*, 2026. URL https://arxiv.org/abs/2608.23862.

Jinsung Yoon, Yanfei Chen, Sercan Arik, and Tomas Pfister. Search-adaptor: Embedding customization for information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10318–10336, 2024b. doi: 10.18653/v1/2024.emnlp-main.576.

HongChien Yu, Chenyan Xiong, and Jamie Callan. Improving query representations for dense retrieval with pseudo relevance feedback. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, 2021.

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*, 2025. URL https://arxiv.org/abs/2506.05176.

Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. BERT-QE: Contextualized query expansion for document re-ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4718–4728, 2020. doi: 10.18653/v1/2020.findings-emnlp.424.

APPENDIX CONTENTS

A Complete FiSA Parameterization 15
A.1 End-to-end retrieval procedure 15
A.2 Network modules and inputs 15
A.3 Teacher targets and losses 16
A.4 Local basis derivation and numerical path 16
A.5 Complexity and deployment invariants 16
B Geometric Interpretation and Limiting Cases 17
B.1 Coordinate dependence and orthogonal rotations 17
B.2 Variance, separation, and stability 17
B.3 Architectural interpretations are not calibrated probabilities 17
B.4 Connections to the ablation variants 17
C Worked Five-Dimensional Example 18
C.1 First-pass neighborhood 18
C.2 Stage 1: coordinate reweighting 18
C.3 Stage 2: local spectrum and Fisher direction 19
C.4 Stage 3: proposal, confidence, and second search 19
D Experimental Protocol 20
D.1 Frozen encoders 20
D.2 Data partition and information flow 21
D.3 Baseline configurations 21
D.4 Seeds and checkpoint selection 21
D.5 Statistical analysis 21
E Full Main Results 22
E.1 Dataset-level secondary metrics 22
E.2 Encoder-level nDCG@10 22
E.3 Cell-level seed dispersion 22
F Ablation Definitions and Full Inference 23
G Efficiency Scope and Integrity Checks 24
G.1 Timing boundary 24
G.2 Result inventory 24
G.3 Claim boundary 24

A COMPLETE FISA PARAMETERIZATION

This appendix expands the three stages in Section 3 into the exact training and inference path used for the released results. The equations follow the frozen implementation; no test labels enter `adapt_query()`.

A.1 END-TO-END RETRIEVAL PROCEDURE

For each query and frozen encoder, inference performs the following operations:

1. Encode and L2-normalize the query q; retrieve the exact Raw Dense top-1000 and retain the first $L = 128$ documents and scores.
2. Predict the coordinate gate g_q and form q_{sel}.
3. Center the top-128 candidates, decompose their sample-space Gram matrix, and recover query-local axes U_q in the original embedding space.
4. Predict positive and negative attention over the 128 candidates.
5. Compute local means, variances, Fisher ratios, learned spectral gates, and the normalized correction direction u_q.
6. Predict residual scale η_q, form q_{prop}, estimate confidence c_q, and obtain q'.
7. Search the complete unchanged index with q' and return the final top-1000.

The first top-128 is not retained as a fixed candidate pool. The second search can introduce any document in the corpus.

A.2 NETWORK MODULES AND INPUTS

Query context. A LayerNorm–Linear–GELU–Dropout block maps the D-dimensional query to a 256-dimensional context. The coordinate predictor and temperature head are linear projections of this context. The coordinate logits, temperature head, residual-scale output layer, and confidence output layer are initialized so that the initial adapter is conservative: the coordinate gate begins at one, the spectral residual begins at zero, and the confidence bias is -2.

Coordinate temperature. The raw temperature is passed through softplus and clamped to $[0.02, 2.0]$. This prevents numerical collapse while preserving meaningful query-specific sharpness. The supervision temperature in Equation 2 is a separate fixed value, $T_{\mathrm{dim}} = 0.01$.

Candidate-attention features. For candidate i, the attention network receives

$$[h_q;\ s_{qi};\ r_{qi};\ (q_{\mathrm{sel}} - \bar{d}_q)^\top (d_{qi} - \bar{d}_q);\ \|d_{qi} - \bar{d}_q\|_2], \tag{11}$$

where r_{qi} decreases linearly from one at rank 1 to zero at rank 128. A shared MLP outputs two logits, one for positive evidence and one for negative evidence, followed by two independent softmax operations over candidates.

Spectral-controller features. For each local axis r, the controller concatenates query context with five axis-specific quantities: standardized log eigenvalue, selected-query coordinate, positive mean, negative mean, and summed within-group variance. The controller returns the gate logit and bounded-correction logit used in Equation 7.

Residual and confidence features. The residual-scale network reads query context and three uncertainty summaries: positive-attention entropy, negative-attention entropy, and normalized entropy of the coordinate-logit softmax. The confidence network instead reads query context, the top-one Raw Dense score, the top-one/top-two margin, and the standard deviation of all 128 base scores. These heads answer different questions: the first controls motion along the spectral direction inside the proposal, while the second controls acceptance of the entire proposal relative to the raw query.

A.3 TEACHER TARGETS AND LOSSES

Graded positive distribution. For relevance label rel_{qi}, the positive target uses gain $g_{qi} = \max(2^{\mathrm{rel}_{qi}} - 1, 0)$ and normalizes it over the candidate set. This target supervises both the positive-attention head and the positive centroid used by the coordinate teacher.

Hard-negative distribution. Candidates with non-positive labels form the negative mask. Their Raw Dense scores are passed through softmax, assigning more mass to negatives that the frozen retriever ranked highly. If the mask is empty, numerical NaNs are replaced by zero. The same distribution supervises negative attention and constructs the negative centroids for coordinate and direction targets.

Coordinate loss. The target in Equation 2 is detached and matched using

$$\mathcal{L}_{\mathrm{dim}} = \mathrm{KL}\big(p_q^{\mathrm{dim}} \,\big\|\, \mathrm{softmax}(z_q)\big)\,. \tag{12}$$

The target is not multiplied directly into the query because it depends on training labels, can contain negative evidence before softmax, and is not bounded as a safe coordinate multiplier.

Direction loss. Label-derived positive and negative weights are applied to the current local coordinates, producing a detached Fisher target that is normalized over local axes. The predicted coefficient vector is also normalized, and $\mathcal{L}_{\mathrm{dir}}$ is one minus their cosine similarity. This target supervises the combined coefficient, not the controller gate and correction separately.

Attention loss. The positive and negative heads use the average of two cross-entropies with the graded positive target and score-weighted negative target. Because the heads are separate distributions, their weights for one document need not sum to one.

Ranking and regularization. The final candidate scores are obtained by dotting the adapted query with the normalized candidate embeddings. The primary ranking loss is cross-entropy against normalized graded gains. The residual penalty is $\mathbb{E}[\eta_q^2]$; the preservation penalty is the mean cosine drift $1 - q'^{\top} q$. The released loss weights are exactly those in Equation 10.

A.4 LOCAL BASIS DERIVATION AND NUMERICAL PATH

Let the compact singular value decomposition of the centered candidate matrix be $X_q = P_q \Sigma_q U_q^{\top}$. Then

$$G_q = \frac{X_q X_q^{\top}}{L-1} = P_q \frac{\Sigma_q^2}{L-1} P_q^{\top}. \tag{13}$$

Thus the eigenvectors V_q in Equation 3 equal P_q, and mapping them through $X_q^{\top}$ recovers the right singular directions U_q. The sample-space decomposition is therefore equivalent to local PCA/SVD while avoiding a $D \times D$ matrix.

All candidate and query vectors are normalized in FP32. The Gram matrix is explicitly symmetrized, eigenvalues are sorted descending and clamped below by 10^{-8}, axes with eigenvalues at most 10^{-7} are zeroed, and recovered basis vectors are normalized. Eigenvectors are defined up to sign, so each axis is multiplied by the sign of its largest-magnitude coordinate. The entire basis construction is detached from gradient computation. On Apple Silicon the detached eigendecomposition runs on CPU because the MPS backend lacks the required kernel; the CUDA and MPS scientific equations are otherwise identical.

A.5 COMPLEXITY AND DEPLOYMENT INVARIANTS

For candidate count L and embedding dimension D, centering and Gram construction cost $O(L^2 D)$, the eigendecomposition costs $O(L^3)$, and mapping the axes back costs $O(L^2 D)$. Temporary storage is $O(LD + L^2)$. With $L = 128$, local adaptation is independent of corpus size; the corpus-dependent overhead is the second index search.

Three implementation boundaries matter for deployment. First, candidate centering and projection are temporary and never overwrite document embeddings. Second, inference consumes only the query, first-pass candidate embeddings, and first-pass scores; test qrels are absent. Third, the output is one normalized query vector, preserving standard maximum-inner-product search and the downstream reader interface.

B Geometric Interpretation and Limiting Cases

B.1 Coordinate dependence and orthogonal rotations

Raw Dense similarity is invariant to a shared orthogonal rotation. If $R^\top R = I$, then $(Rq)^\top (Rd) = q^\top d$. The first FiSA stage is not rotation invariant because it predicts a diagonal gate in the frozen encoder basis. After rotation, the same diagonal gate no longer represents the same linear operator. This is why each frozen encoder receives its own adapter and why we do not interpret individual coordinate indices as universal semantic factors. The local spectral stage is different: it reconstructs axes from the candidate neighborhood after the encoder has been fixed, so its axes adapt to the rotated candidate cloud.

This distinction gives FiSA two complementary forms of adaptation. The coordinate gate exploits stable regularities learned in one encoder space, while the local basis recovers directions that can mix many encoder coordinates for one query. A coordinate-only method can be effective when the encoder basis already exposes the relevant distinction. A local correction is most useful when separation lies along a combination of coordinates that is not naturally axis aligned.

B.2 Variance, separation, and stability

The local eigenvalue λ_{qr} measures how much the retrieved candidates vary along axis r. It says nothing about which side contains relevant evidence. Two failure modes illustrate why variance alone is insufficient. First, an axis can have high variance because the candidates mix writing styles or document lengths while positive and negative evidence overlap. Second, an axis can have modest variance but nearly perfect class separation, as in the first axis of the worked example. FiSA therefore uses the spectrum as a description of available local directions and the Fisher ratio as a relevance-sensitive statistic.

The denominator in Equation 6 is equally important. A large positive–negative mean difference is not reliable if each evidence group is widely dispersed. Conversely, a smaller mean gap can be useful when both groups are compact. The diagonal approximation evaluates this trade-off one local axis at a time. The learned controller then handles cases where the empirical ratio is numerically extreme or systematically biased.

B.3 Architectural interpretations are not calibrated probabilities

The spectral gate, residual scale, and adaptation confidence have useful functional interpretations, but they are not independently supervised or probability calibrated. The spectral gate and bounded correction are produced by the same controller and can partially compensate for one another. Likewise, c_q is trained through ranking and preservation objectives rather than a binary label that states whether adaptation is correct. We therefore describe these quantities as shrinkage, residual, and confidence controls in the architectural sense. Their values should not be read as calibrated probabilities of relevance or success.

B.4 Connections to the ablation variants

Several limiting configurations recover the controlled variants in Table 4. Setting $g_q = \mathbf{1}$ removes query-aware coordinate selection. Removing u_q and retaining the coordinate and confidence heads gives Dimension Only. Replacing the learned attention distributions with fixed softmaxes of positive and negative base scores gives Fixed PRF Attention. Replacing U_q with one source-corpus PCA basis gives Global Basis. Finally, setting $c_q = 1$ removes the trust region and always accepts the proposal. These connections make the ablation matrix a decomposition of one architecture rather than a collection of unrelated alternatives.

C WORKED FIVE-DIMENSIONAL EXAMPLE

This section gives a concrete numerical example of all three stages. It compresses the released top-128, high-dimensional setting to ten candidates in five dimensions while preserving the operations. Values attributed to the local basis and Fisher statistics are computed from the example documents; controller and confidence outputs are illustrative network predictions.

C.1 FIRST-PASS NEIGHBORHOOD

Consider the normalized query

$$q = (0.50, 0.50, 0.50, 0.40, 0.30). \tag{14}$$

The first ranking contains two relevant documents followed by eight hard negatives. The third and fifth coordinates are active in the query and also large for many negatives, allowing those negatives to receive high Raw Dense scores.

Table A1: Ten-document, five-dimensional first-pass example. Labels are available only during training.

Doc.	Vector	$q^\top d_i$	Label
d_1	$(.66, .61, .10, .42, .05)$	.868	1
d_2	$(.61, .66, .12, .41, .06)$	.877	1
d_3	$(.12, .20, .88, .08, .40)$	.752	0
d_4	$(.20, .12, .86, .10, .42)$	.756	0
d_5	$(.15, .55, .70, -.10, .40)$	.780	0
d_6	$(.55, .15, .68, -.15, .42)$	.756	0
d_7	$(.10, .65, .60, -.20, .40)$	.715	0
d_8	$(.65, .10, .58, -.20, .43)$	.714	0
d_9	$(.25, .25, .75, -.25, .50)$	.675	0
d_{10}	$(.30, .30, .72, -.28, .47)$	.689	0

C.2 STAGE 1: COORDINATE REWEIGHTING

The positive centroid is

$$\mu^+ = (0.635, 0.635, 0.110, 0.415, 0.055), \tag{15}$$

and score-weighted negatives give approximately

$$\mu^- = (0.289, 0.290, 0.722, -0.122, 0.429). \tag{16}$$

Their difference and the coordinate teacher are

$$\mu^+ - \mu^- \approx (0.346, 0.345, -0.612, 0.537, -0.374), \tag{17}$$

$$t = q \odot (\mu^+ - \mu^-) \approx (0.173, 0.172, -0.306, 0.215, -0.112). \tag{18}$$

With $T_{\mathrm{dim}} = 0.01$, the target distribution concentrates on the fourth coordinate, with smaller mass on the first two and nearly zero mass on the third and fifth.

Suppose the trained query-only predictor outputs $z = (0.8, 0.7, -1.0, 1.2, -0.7)$ and $\tau(q) = 0.5$. Since $\bar{z} = 0.2$,

$$g = 2\sigma((z - \bar{z})/0.5) \approx (1.537, 1.462, 0.166, 1.762, 0.284). \tag{19}$$

After multiplying and normalizing,

$$\boxed{q_{\mathrm{sel}} \approx (0.601, 0.572, 0.065, 0.551, 0.067)}. \tag{20}$$

The representation remains five-dimensional: the stage changes relative coordinate contributions rather than truncating the vector.

C.3 Stage 2: local spectrum and Fisher direction

The candidate centroid is approximately

$$\bar{d} = (0.360, 0.360, 0.602, -0.017, 0.356). \tag{21}$$

Centering and decomposing the ten candidates gives the following five local axes and eigenvalues:

$$u_1 = (-.350, -.355, .609, -.499, .363), \qquad \lambda_1 = .194, \tag{22}$$
$$u_2 = (.724, -.683, -.058, -.064, .042), \qquad \lambda_2 = .053, \tag{23}$$
$$u_3 = (-.262, -.391, .354, .797, -.135), \qquad \lambda_3 = .031, \tag{24}$$
$$u_4 = (.522, .482, .695, .086, -.073), \qquad \lambda_4 = .001, \tag{25}$$
$$u_5 = (.108, .152, -.131, .324, .918), \qquad \lambda_5 \approx 0. \tag{26}$$

The axes are mixtures of the original coordinates. They describe directions of candidate variation, not dimensions selected by a global PCA model.

Table A2: Candidate coordinates in the local spectral basis.

Doc.	u_1	u_2	u_3	u_4	u_5
d_1	-.829	.035	.035	-.010	-.003
d_2	-.809	-.036	.027	.001	.003
d_3	.279	-.085	.296	-.002	-.015
d_4	.269	.028	.314	-.007	.012
d_5	.126	-.283	-.057	.046	.009
d_6	.148	.290	-.055	.041	-.004
d_7	.096	-.374	-.198	-.014	-.002
d_8	.097	.405	-.138	-.008	.004
d_9	.336	.008	-.080	-.038	.008
d_{10}	.288	.012	-.144	-.008	-.012

The first axis provides a clear example of relevance separation: both positives have coordinate about -0.82, while every negative lies on the positive side. Assume the learned attention is close to an idealized distribution, with $a^+ \approx (0.5, 0.5, 0, \ldots, 0)$ and negative mass spread over the eight hard negatives. The resulting Fisher statistics are shown below.

Table A3: Example positive/negative moments and Fisher-style coefficients.

Axis	μ^+	μ^-	v^+	v^-	f
u_1	-.819	.204	≈ 0	.008	-109.84
u_2	≈ 0	-.001	.001	.060	.015
u_3	.031	-.004	≈ 0	.035	.981
u_4	-.005	.002	≈ 0	$\approx .001$	-3.880
u_5	≈ 0	≈ 0	≈ 0	≈ 0	-.506

The extreme magnitude on u_1 illustrates why a learned controller is needed. Suppose it outputs gates $m = (.02, .15, .80, .10, .05)$ after sigmoid and bounded corrections $\delta = (.50, -.10, .30, .50, .20)$. Then

$$\alpha = m \odot (f + \delta) \approx (-2.187, -.013, 1.025, -.338, -.015). \tag{27}$$

Combining the axes and normalizing yields

$$\boxed{u \approx (0.127, 0.090, -0.492, 0.768, -0.378)}. \tag{28}$$

The controller has strongly shrunk the unstable first Fisher ratio while retaining its useful orientation and combining it with other axes.

C.4 Stage 3: proposal, confidence, and second search

Assume the residual network predicts $\eta = 0.35$. The proposal is

$$q_{\text{prop}} = \text{Norm}(q_{\text{sel}} + 0.35u) \tag{29}$$
$$\approx (0.533, 0.498, -0.089, 0.677, -0.054). \tag{30}$$

This proposal aggressively reverses the third and fifth coordinates. If the Raw Dense top score is .880, the top-one/top-two margin is .010, and score standard deviation is .065, suppose the confidence network predicts $c = 0.60$. Interpolation with the raw query gives

$$\boxed{q' \approx (0.558, 0.535, 0.158, 0.607, 0.094)}. \tag{31}$$

The confidence gate accepts the direction of the proposal while avoiding its sign reversals in the final normalized query.

Table A4: Raw and adapted scores for the example candidates. The released method would search the complete corpus, not only these ten documents.

Doc.	Label	Raw score	Adapted score
d_1	1	.869	.972
d_2	1	.880	.970
d_3	0	.754	.400
d_4	0	.764	.416
d_5	0	.786	.469
d_6	0	.761	.446
d_7	0	.718	.416
d_8	0	.716	.428
d_9	0	.675	.287
d_{10}	0	.690	.316

The example separates the roles of the three stages. Coordinate selection suppresses the influence of predicted query-side nuisance coordinates without changing the embedding width; the Fisher-spectral stage identifies a candidate-local direction in temporary local coordinates and maps it back to the original embedding space; and the residual and confidence gates decide how much of that direction can be accepted safely.

D Experimental Protocol

D.1 Frozen encoders

Table A5: Frozen embedding models and encoding rules. Revisions are the exact commit hashes used by the formal configuration.

Alias	Model / revision	Pooling	Prefixes
BGE-base	`BAAI/bge-base-en-v1.5` `a5beb1e3e68b9ab74eb54cfd186867f64f240e1a`	CLS	Query: “Represent this sentence for searching relevant passages: ”
E5-base	`intfloat/e5-base-v2` `f52bf8ec8c7124536f0efb74aca902b2995e5bcd`	Mean	Query: “query: ”; document: “passage: ”
BGE-large	`BAAI/bge-large-en-v1.5` `d4aa6901d3a41ba39fb536a557fa166f842b0e09`	CLS	Query: “Represent this sentence for searching relevant passages: ”
Qwen3-0.6B	`Qwen/Qwen3-Embedding-0.6B` `97b0c614be4d77ee51c0cef4e5f07c00f9eb65b3`	Last token	Instruction-form web-search query prefix from the model card
BGE-M3	`BAAI/bge-m3` `5617a9f61b028005a4858fdac845db406aefb181`	CLS	No additional query or document prefix

All encoders are frozen. Embeddings are L2-normalized before indexing and scoring. The formal encoding batch sizes are 64 for BGE-base and E5-base, 16 for BGE-large, and 8 for Qwen3-0.6B and BGE-M3. Query batches contain 128 items and document search is chunked into blocks of 50,000 documents.

D.2 Data partition and information flow

The official training queries of NFCorpus, SciFact, and FiQA are assigned to training or development with a stable query-ID hash; 10% enter development. The training loader keeps only queries whose Raw Dense top-128 contains at least one judged relevant document. Development and test evaluation retain complete query sets, including queries with no positive in the feedback neighborhood. ArguAna is never used for training, development selection, hyperparameter tuning, or global-basis fitting.

Each encoder receives one independently trained FiSA adapter. Source-domain training is balanced across the three datasets so that a larger query set does not dominate minibatch sampling. Formal training is deterministic, uses exponential moving average with decay .99, and runs for at most 30 epochs with patience 7. The optimizer learning rate is 10^{-4}, weight decay is .01, batch size is 8, hidden width is 256, and dropout is .1.

D.3 Baseline configurations

Table A6: Frozen baseline and FiSA settings.

Method	Formal configuration
Raw Dense	One exact full-corpus search with the frozen normalized query and document embeddings; top-1000 output.
SpecTemp-256	Target dimension 256; corpus covariance sample up to 1,000,000 documents; deterministic seed 1999; transformed query and document index.
ECLIPSE-PRF	$k_+ = 1$; $k_- \in \{10, 50, 100\}$; positive weight 1.0; negative weight $\in \{0.5, 1.0\}$; retained ratio $\in \{0.25, 0.50, 0.75\}$; development-selected per encoder.
FiSA	Candidate count 128; Fisher ridge 10^{-3}; hidden width 256; dropout .1; 20 seeds; full-corpus second search.

SpecTemp requires a transformed document index for the selected width. ECLIPSE-PRF and FiSA reuse the Raw Dense document index and first-pass results. ECLIPSE-PRF constructs a signed importance score in the original encoder basis, masks coordinates according to the selected ratio, and performs a second search.

D.4 Seeds and checkpoint selection

The main seeds are

$$\{13, 21, 42, 87, 100, 7, 17, 29, 37, 53, 61, 73, 89, 101, 113, 127, 139, 151, 163, 179\}. \quad (32)$$

The original five seeds $\{13, 21, 42, 87, 100\}$ are used for every ablation. Checkpoint selection evaluates full-corpus development nDCG@10 on all three source domains. Feasible checkpoints, defined as non-negative delta from Raw Dense on each development domain, always outrank infeasible checkpoints. Within the feasible set, macro development nDCG@10 selects the epoch. If no epoch is feasible, the fixed fallback chooses the checkpoint with the largest worst-domain delta. This rule is applied without test-label access.

D.5 Statistical analysis

The macro score weights each of the 20 encoder–dataset cells equally. For learned models, cell scores first average physical seeds. Hierarchical bootstrap resamples query clusters shared across methods and encoders and independently resamples the 20 learned seeds for each encoder. Randomization flips paired query-cluster differences and uses 10,000 draws; Holm adjustment is applied across the three main baseline comparisons and across the five ablation comparisons. The resulting intervals are conditional on the named datasets, encoders, and benchmark queries; they are not uncertainty estimates over unseen task populations.

E FULL MAIN RESULTS

Tables 2 and 3 in the main text report macro effectiveness and paired nDCG@10 inference. This appendix provides the detailed dataset, encoder, and seed breakdowns.

E.1 DATASET-LEVEL SECONDARY METRICS

Table A7: Dataset-level effectiveness averaged over five encoders. FISA averages 20 seeds.

Dataset	Method	nDCG@10	MRR@10	R@100	R@1000
NFCORPUS	Raw Dense	.3592	.5591	.3285	.6472
	SPECTEMP-256	.3421	.5297	.3158	.6237
	ECLIPSE-PRF	.3705	.5493	.3484	.6648
	FISA	**.3863**	**.5693**	**.3610**	**.6775**
SCIFACT	Raw Dense	.7104	.6752	.9469	.9927
	SPECTEMP-256	.6974	.6605	.9432	.9907
	ECLIPSE-PRF	.7044	.6668	.9495	**.9953**
	FISA	**.8248**	**.8022**	**.9552**	.9918
FIQA	Raw Dense	.4276	.5108	.7516	.9179
	SPECTEMP-256	.4067	.4879	.7266	.9082
	ECLIPSE-PRF	.4252	.4967	.7561	.9193
	FISA	**.4393**	**.5134**	**.7664**	**.9212**
ARGUANA	Raw Dense	.1759	.0776	.9484	.9932
	SPECTEMP-256	.1713	.0752	.9538	.9932
	ECLIPSE-PRF	**.1934**	**.0852**	**.9804**	**.9960**
	FISA	.1884	.0830	.9645	.9952

The table above is intended as a compact appendix summary; authoritative values are the machine-readable aggregate files in the artifact. The primary conclusions do not rely on rounding in this display.

E.2 ENCODER-LEVEL NDCG@10

Table A8: Mean nDCG@10 by encoder, averaged over four datasets.

Encoder	Raw Dense	SPECTEMP-256	ECLIPSE-PRF	FISA
BGE-base	.4288	.4125	.4318	**.4636**
E5-base	.3986	.3787	.4105	**.4488**
BGE-large	.4451	.4291	.4515	**.4774**
Qwen3-0.6B	.4371	.4295	.4412	**.4760**
BGE-M3	.3818	.3722	.3820	**.4327**

E.3 CELL-LEVEL SEED DISPERSION

The largest cell-level seed standard deviations remain below .0055 nDCG@10. Variation is higher on SCIFACT, NFCORPUS, and FIQA for some encoders than on ARGUANA, but every cell-level range is small relative to the main improvements over Raw Dense. This table also shows why a 20-seed average is informative even when the macro standard deviation is very small: local cells can vary more than the aggregate after positive and negative deviations cancel.

Table A9: FiSA nDCG@10 seed dispersion for every encoder–dataset cell. The range reports the minimum and maximum over 20 physical training seeds.

Encoder	Dataset	Mean	Seed SD	Range
BGE-base	NFCorpus	0.3979	0.0039	[0.3884, 0.4027]
	SciFact	0.8465	0.0043	[0.8351, 0.8534]
	FiQA	0.4127	0.0018	[0.4089, 0.4157]
	ArguAna	0.1974	0.0015	[0.1947, 0.2006]
E5-base	NFCorpus	0.3890	0.0029	[0.3841, 0.3953]
	SciFact	0.8308	0.0046	[0.8241, 0.8383]
	FiQA	0.4216	0.0043	[0.4138, 0.4296]
	ArguAna	0.1539	0.0032	[0.1486, 0.1593]
BGE-large	NFCorpus	0.4088	0.0042	[0.3995, 0.4154]
	SciFact	0.8348	0.0041	[0.8248, 0.8407]
	FiQA	0.4609	0.0036	[0.4520, 0.4651]
	ArguAna	0.2053	0.0014	[0.2032, 0.2075]
Qwen3-0.6B	NFCorpus	0.3887	0.0046	[0.3764, 0.3976]
	SciFact	0.8228	0.0030	[0.8194, 0.8283]
	FiQA	0.4727	0.0046	[0.4580, 0.4779]
	ArguAna	0.2197	0.0014	[0.2172, 0.2225]
BGE-M3	NFCorpus	0.3473	0.0026	[0.3432, 0.3523]
	SciFact	0.7889	0.0055	[0.7808, 0.7980]
	FiQA	0.4289	0.0034	[0.4204, 0.4351]
	ArguAna	0.1657	0.0014	[0.1637, 0.1691]

F Ablation Definitions and Full Inference

All variants inherit the Full FiSA optimizer, balanced sampling, EMA, data partition, candidate count, checkpoint rule, and second full-corpus evaluation. They are defined as follows.

w/o Coordinate Reweighting. Fix $g_q = \mathbf{1}$ and remove $\mathcal{L}_{\mathrm{dim}}$; all candidate-local and confidence components remain.

Fixed PRF Attention. Replace learned evidence with $a_q^+ = \mathrm{softmax}(s_q)$ and $a_q^- = \mathrm{softmax}(-s_q)$; remove the attention loss.

Dimension Only. Retain the coordinate and confidence gates but remove candidate attention, local basis construction, Fisher correction, and their losses.

Global Basis. Replace U_q with an encoder-specific rank-128 PCA basis estimated once from 8,192 source-domain documents using seed 2027. ArguAna documents and all qrels are excluded from fitting; all other FiSA components remain.

w/o Trust Region. Fix $c_q = 1$ and remove the preservation loss, forcing every proposal to be accepted.

Table A10: Full paired ablation means and seed standard deviations in nDCG@10.

Variant	NFCorpus	SciFact	FiQA	ArguAna	Macro
Full FiSA	.3857±.0028	.8237±.0012	.4395±.0007	.1885±.0008	.4593±.0008
w/o Coordinate Reweighting	.3729±.0012	.7169±.0009	.4360±.0004	.1919±.0007	.4294±.0006
Fixed PRF Attention	.3729±.0012	.8074±.0013	.4353±.0010	.1741±.0018	.4474±.0005
Dimension Only	.3687±.0010	.8160±.0037	.4324±.0006	.1744±.0017	.4479±.0012
Global Basis	.3836±.0013	.8197±.0017	.4339±.0012	.1942±.0009	.4579±.0004
w/o Trust Region	.3737±.0042	.8013±.0055	.4276±.0014	.1830±.0017	.4464±.0023

The Global Basis comparison is intentionally described as mixed. Its adjusted randomization value is below .05, but the hierarchical confidence interval includes zero, Full fails to win eight cells, and

Table A11: Paired primary-metric inference for the ablations. Δ is Full minus variant macro nDCG@10. The final column counts cells where Full does not win.

Variant	Δ	95% CI	Holm p	Non-win cells
w/o Coordinate Reweighting	.02991	[.02366,.03641]	.00050	3/20
Fixed PRF Attention	.01193	[.01000,.01392]	.00050	0/20
Dimension Only	.01146	[.00922,.01380]	.00050	0/20
Global Basis	.00149	[−.00029,.00327]	.02950	8/20
w/o Trust Region	.01295	[.01053,.01559]	.00050	0/20

Global Basis is stronger on the unseen domain. The two procedures answer different conditional questions, and neither justifies universal local-basis superiority.

G Efficiency Scope and Integrity Checks

G.1 Timing boundary

The release includes method-runtime measurements for the ablation variants, but the recorded scope excludes shared encoding, indexing, and cached first-pass candidate mining. It is therefore useful for checking that variants execute and for diagnosing local overhead, not for comparing end-to-end deployment latency. Peak GPU memory is recorded as zero for the macOS runs because the platform did not expose a CUDA-style peak-memory counter. We keep these fields in the artifact for provenance and avoid converting them into unsupported performance claims.

G.2 Result inventory

The final main inventory contains $5 \times 4 \times 20 = 400$ physical FISA results and $3 \times 5 \times 4 = 60$ deterministic baseline results. The ablation inventory adds 5 variants $\times 5$ encoders $\times 4$ datasets $\times 5$ seeds $= 500$ physical results, paired with the matching five Full FISA seeds. Every aggregate was recomputed from saved per-query records.

The audit found no missing or duplicated experiment identities, NaNs, out-of-range metrics, or query-count mismatches. Candidate-cache and query identities match across compared methods. The original five Full seeds and the 15-seed extension share the same frozen scientific source hash, and the ablation variants were frozen together before official-test evaluation. Deterministic baseline outputs are physically stored once and are never replicated to create artificial seed variance.

G.3 Claim boundary

The artifact supports claims about retrieval effectiveness on the named English benchmarks and frozen embedding models. It does not evaluate generated answers, faithfulness, citation correctness, multilingual retrieval, or reader behavior. The main comparison also contrasts a supervised source-trained adapter with baselines that do not use equivalent source qrels. Component causality is therefore established most directly by the paired ablations, which share supervision, seeds, data partitions, and checkpoint selection.