FAILURE IS NOT ABSENCE: SELECTIVE LOW RANK SUPERVISION FOR DENSE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense retrieval failures are often attributed to representations that do not sufficiently capture the discriminative structure required to distinguish relevant from non relevant documents. We challenge this view by showing that retrieval failure does not necessarily imply the absence of such structure. Using frozen single vector representations, we show that queries that fail under the original retrieval function are recovered through low rank readouts. We further find that the discriminative signals enabling this recovery are extracted even from very low dimensional subspaces. By decomposing the full space score into selected subspace and residual contributions, we show that the low rank contributions associated with recovery are counteracted by residual contributions, leading to dense retrieval failure.

Building on these findings, we propose Selective Low Rank Supervision, which selects beneficial rank one signals and aligns the full embedding representation with the selected signals during training. The proposed method requires no additional subspace computation at inference, while outperforming standard contrastive fine tuning. These results indicate that the low rank structure revealed by our analysis is not only informative for understanding retrieval failure, but can also provide an effective training signal for improving the full embedding representation for retrieval.

1 INTRODUCTION

Dense retrieval represents queries and documents as dense vectors and retrieves documents according to their similarity in a shared embedding space (Karpukhin et al., 2020). Despite its effectiveness, recent studies have identified several limitations of single vector retrieval. The retrieval outcomes expressible by single vector embeddings are fundamentally constrained by the embedding dimension (Weller et al., 2026). Dense embeddings can also fail to capture detailed entity and event information while preserving overall semantics (Xu et al., 2025). These studies point to limitations in what a single dense representation can express.

Other studies show that retrieval performance also depends on how the learned representation interacts with the similarity function. Cosine and Euclidean similarity can fail to reflect semantic relationships when dense representations lie on nonlinear manifolds (Liu et al., 2025). The geometric structure of the embedding space, particularly its isotropy, can also affect retrieval based on similarity (Kim et al., 2024). Taken together, these studies suggest that the retrieval outcome alone does not clearly reveal whether a useful retrieval component remains in the representation. This uncertainty leads to the central question of our work: does dense retrieval failure necessarily imply the absence of useful discriminative structure?

To answer this question, we analyze the pretrained representation through subspaces while keeping it fixed. The study follows two directions. One analyzes the properties of recovery subspaces across different ranks, including their recoverability, influence, discriminative concentration, and regularity. The other investigates whether these findings can be used as a learning signal to improve the full embedding representation.

Building on these analyses, we propose Selective Low Rank Supervision, which selects a beneficial recovery subspace and incorporates its contribution into the training target distribution. Rather than

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.

directly supervising this subspace, the method incorporates its contribution only when it improves the relative preference for the relevant document. This supervision is used only during training, while inference remains unchanged and requires no additional subspace computation or projection.

Our main contributions are as follows:

We observe that failed queries can be recovered from low rank subspaces of fixed pretrained representations, and further analyze why retrieval can still fail even when such recovery signals are present.

We characterize recovery subspaces in terms of their recoverability, the influence of components outside the recovery subspace, and their discriminative concentration and regularity across different ranks. These characteristics allow us to examine whether recovery subspaces can be used as learning signals for improving the full embedding representation.

We propose Selective Low Rank Supervision, which selects a beneficial rank 1 recovery subspace and incorporates its contribution into the training target distribution. The proposed supervision can also be applied with other sampling strategies.

2 Related Work

2.1 Understanding Dense Retrieval Failure

Recent work has focused on several sources of failure in dense retrieval. Single vector retrieval is fundamentally constrained by embedding dimensionality (Weller et al., 2026), while dense representations can lose detailed entity and event information required for matching (Xu et al., 2025). Document embeddings can also exhibit positional bias in how information is represented (Coelho et al., 2024), and representing a document with a single vector can limit the modeling of diverse query–document relationships (Kang et al., 2025). Beyond representation capacity, retrieval is affected by embedding isotropy (Kim et al., 2024) and nonlinear manifold structure (Liu et al., 2025), which influence how similarity reflects relevance. Information within text embeddings can also be unevenly organized across directions, with substantial changes concentrated in a principal component (Nie et al., 2025). However, these studies leave it unclear whether the structures revealed by these analyses can actually recover failed retrievals. We address this by analyzing recovery within fixed pretrained subspaces and the effects of the remaining components.

2.2 Learning Dense Representations for Retrieval

Dense representation learning has explored richer forms of supervision for retrieval. M3-Embedding integrates relevance signals from different retrieval functions through self knowledge distillation (Chen et al., 2024), while PairDistill transfers pairwise relevance information from a reranker to a dense retriever (Huang & Chen, 2024). Fine grained distillation introduces supervision across multiple representation granularities while preserving the conventional dense retrieval setting (Zhou et al., 2024). More recent embedding distillation methods transfer internal relational structure (Truong et al., 2025) or align student representations with a stronger teacher embedding space (Vujanic & Rückstieß, 2026). In contrast, Selective Low Rank Supervision derives its additional signal from recovery directions within the pretrained representation itself and selectively incorporates beneficial recovery contributions into the target distribution.

3 Analyzing Dense Retrieval Failure

3.1 Problem Setup and Analysis

Let $q \in \mathbb{R}^D$, $d^+ \in \mathbb{R}^D$, and $d^- \in \mathbb{R}^D$ denote the query, relevant document, and non relevant document embeddings obtained from the embedding model $f(\cdot)$, respectively, where D is the full embedding dimension of the model. Before retrieval, all query and document embeddings are L2-normalized. Therefore, cosine similarity is equivalent to the dot product, and we define the retrieval score between a query q and a document d as

$$s(q, d) = q^\top d. \tag{1}$$

Dense retrieval orders documents according to $s(q, d)$. When d^+ is not included in the top-k results, the retrieval is considered a failure. However, we do not assume that such a failure necessarily means that the information required for correct retrieval is absent from the embedding.

To examine this possibility, we select r orthonormal directions and represent them by U, where $1 \leq r \leq D$. Here, r denotes the number of selected directions, while D denotes the full embedding dimension.

Let $z \in \mathbb{R}^D$ denote either a query or document embedding. The coordinates of z along the selected directions are given by $U^\top z$. Since retrieval is performed in the original D-dimensional embedding space, we map these coordinates back through U and define the corresponding projection as

$$U(U^\top z) = UU^\top z, \qquad P = UU^\top, \qquad z_U = Pz. \tag{2}$$

Here, P denotes the projection matrix onto the subspace spanned by U, and z_U is the projected representation. All embeddings are kept fixed so that any change in retrieval results comes only from the selected subspace rather than from changes in the encoder.

3.2 Retrieval Recoverability

To analyze whether partial subspaces U can recover retrieval failures, we define a Failure-to-Success (F2S) case as a query that fails in the original embedding space but succeeds after projection onto the selected subspace.

$$\text{F2S} : \text{Failure}(q, d^+) \rightarrow \text{Success}(Pq, Pd^+). \tag{3}$$

The main analysis in this section uses the ArguAna dataset with the Qwen model. For each rank, we report Unique F2S, which counts the failed queries recovered by at least one candidate subspace constructed with that rank. We also report Cumulative Success, which counts the successful queries accumulated from the cosine baseline and the recovered F2S cases up to the corresponding rank.

Table 1: Cumulative retrieval recovery on ArguAna using Qwen.

Method	Unique F2S@10	Cumulative Success@10
Cosine baseline	–	1,298 / 1,401 (92.65%)
Cosine ∪ Rank 1	68	1,366 / 1,401 (97.50%)
Cosine ∪ Rank 2	89	1,387 / 1,401 (99.00%)
Cosine ∪ Rank 4	99	1,397 / 1,401 (99.71%)
Cosine ∪ Rank 8	102	1,400 / 1,401 (99.93%)
Cosine ∪ Rank 16	103	1,401 / 1,401 (100.00%)
Cosine ∪ Rank 32	103	1,401 / 1,401 (100.00%)
Cosine ∪ All Ranks	103	1,401 / 1,401 (100.00%)

As shown in Table 1, failed queries can be recovered using only selected directions of the frozen embedding. A large number of failures are recovered even with a small number of selected directions, and combining all tested ranks covers all queries in the analysis. These results show that recoverable components exist in the frozen representation.

3.3 Effect of Non-Recovery Components

Section 3.2 shows that some failed queries can be recovered from a selected subspace U, but this observation alone does not explain why the same queries fail when the full representation is used. We therefore examine whether components outside the recovery subspace interfere with correct retrieval.

Since U is composed of orthonormal directions, the projection matrix P divides an embedding into the recovery component and the remaining component as

$$z = Pz + (I - P)z, \qquad s_\alpha(q, d) = q^\top Pd + (1 - \alpha)q^\top (I - P)d. \tag{4}$$

Here, α controls how strongly the remaining component is suppressed. When $\alpha = 0$, the score is identical to the original full space score, while larger values of α reduce the contribution from components outside the recovery subspace.

We evaluate whether increasing α changes F2S queries from failure to success. We also report the maximum number of Success to Failure (S2F) cases observed during suppression to check whether this intervention disrupts queries that were already successfully retrieved.

Table 2: Retrieval changes after suppressing the non-recovery component.

Rank	F2S Queries	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.9$	Max. S2F
1	68	1	3	24	0
2	89	3	7	46	0
4	99	5	10	50	0
8	102	7	14	63	0
16	103	6	17	70	0
32	103	12	23	84	0

As shown in Table 2, more F2S queries are recovered as α increases. These results indicate that some components outside the recovery subspace can contribute to retrieval failure by interfering with the recovery signal. At the same time, almost no S2F cases are observed during suppression, with only one case at Rank 16.

3.4 Regularity and Concentration of Recovery Subspaces

We next analyze whether recovery subspaces have properties that can be used as learning targets and whether they contain more discriminative information than expected from their dimensionality.

For queries that can be recovered by multiple subspaces of the same rank, we compare how much the successful recovery subspaces overlap relative to random subspaces of the same dimensionality. Let $U^{(a)}$ and $U^{(b)}$ denote two successful rank-r recovery subspaces found for the same query. We define their overlap as

$$\mathrm{Overlap}\left(U^{(a)}, U^{(b)}\right) = \frac{1}{r}\left\|\left(U^{(a)}\right)^{\top} U^{(b)}\right\|_F^2. \tag{5}$$

The overlap is averaged over successful recovery-subspace pairs and compared with the corresponding random baseline. We also measure how much of the difference between the relevant and non-relevant document representations is captured by the recovery subspace. Let P denote the projection matrix of a rank-r recovery subspace. We define the discriminative proportion, its dimensional expectation, and enrichment as

$$\mathrm{Disc}(r) = \frac{\|P(d^+ - d^-)\|_2^2}{\|d^+ - d^-\|_2^2}, \qquad \mathrm{Expected}(r) = \frac{r}{D}, \qquad \mathrm{Enrichment}(r) = \frac{\mathrm{Disc}(r)}{\mathrm{Expected}(r)} \tag{6}$$

Table 3: Overlap and discriminative concentration of recovery subspaces.

Rank	$N_{\geq 2}$	Success-U	Random	Ratio	Disc. (%)	Expected (%)	Enrich.
1	37	0.001181	0.000977	**1.209**	0.324	0.098	**3.317**
2	65	0.001903	0.001952	0.975	0.289	0.195	1.480
4	93	0.003891	0.003906	0.996	0.346	0.391	0.885
8	101	0.007891	0.007815	1.010	0.581	0.781	0.744
16	102	0.015604	0.015624	0.999	1.143	1.562	0.732
32	103	0.031252	0.031251	1.000	2.430	3.125	0.778

The discriminative and expected proportions are reported as percentages in Table 3, where Rank 1 shows the clearest difference from the random baseline with an overlap ratio of 1.321. As the rank increases, the overlap becomes nearly indistinguishable from the random baseline.

More notably, the discriminative concentration at Rank 1 is substantially higher than its dimensional expectation. Although a Rank 1 subspace accounts for only 0.130% of the full embedding dimensions, it captures 0.412% of the positive-negative difference, corresponding to an enrichment of

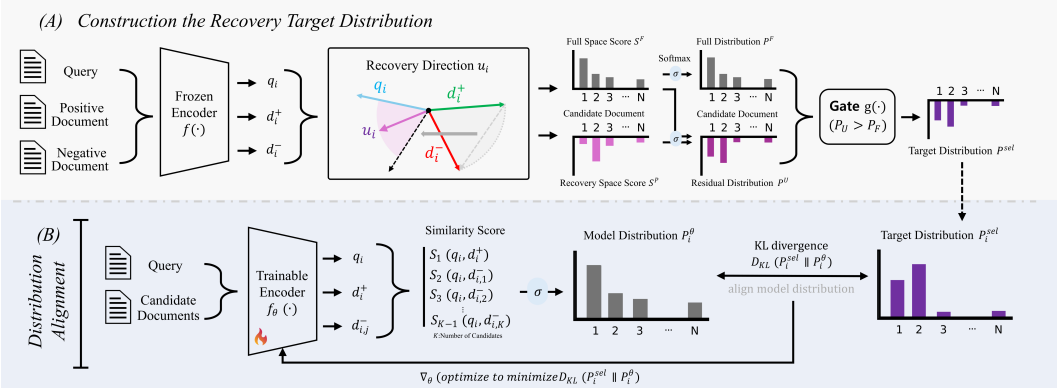

Figure 1: Overview of Selective Low Rank Supervision. (A) A frozen pretrained encoder is used to construct a query specific Rank 1 recovery direction and two candidate distributions from the full space and recovery augmented scores. The recovery contribution is selected only when it increases the relative preference for the positive document, producing the target distribution p_i^{sel}. (B) A trainable encoder is optimized by aligning its model distribution p_i^θ with the selected target distribution using KL divergence.

3.15×. Both effects decrease as the rank increases. These findings motivate the use of the recovery direction U in our learning design.

4 Learning from Recovery Directions

Based on the analysis in Section 3, Rank 1 shows the clearest structure among the tested recovery subspaces, but its directions still exhibit substantial randomness. We therefore derive the recovery direction U directly from the relationship among the query, positive document, and negative document, rather than searching for an arbitrary subspace.

This direction is used to construct a training target distribution that reflects beneficial recovery contributions.

4.1 Deriving the Recovery Direction

For each training query, let q_i^0, d_i^{+0}, and d_i^{-0} denote the L2-normalized embeddings produced by the pretrained encoder. Here, d_i^{+0} is the relevant document and d_i^{-0} is a non-relevant document. We first define the positive-negative separation direction δ_i and combine it with the query representation to obtain the recovery direction u_i:

$$\delta_i = \frac{d_i^{+0} - d_i^{-0}}{\left\| d_i^{+0} - d_i^{-0} \right\|_2}, \qquad u_i = \frac{q_i^0 + \delta_i}{\left\| q_i^0 + \delta_i \right\|_2}. \tag{7}$$

By the Cauchy–Schwarz inequality, δ_i gives the unit direction that maximizes the separation between d_i^{+0} and d_i^{-0}. Since δ_i does not yet incorporate the query representation, u_i combines this separation direction with q_i^0. Thus, u_i jointly reflects the query and the separation between its positive and negative documents. The direction u_i is fixed during optimization.

4.2 Selective Target Distribution

The constructed direction does not necessarily improve retrieval for every training example. We therefore use it only when its contribution improves the relative score distribution of the relevant document.

For each query q_i, let $C_i = \{d_{i,0}^0, d_{i,1}^0, \ldots, d_{i,|C_i|-1}^0\}$ denote its training candidate document set. We index the relevant document as $d_{i,0}^0 = d_i^{+0}$, while the remaining documents $d_{i,j}^0$ for $j \geq 1$ are

non-relevant candidates. For each candidate $d^0_{i,j} \in C_i$, we compute the full space score and the Rank 1 projected score as

$$s^0_{F,ij} = q_i^{0\top} d^0_{i,j}, \qquad s^0_{P,ij} = (q_i^{0\top} u_i)(u_i^\top d^0_{i,j}). \tag{8}$$

To compare the relative retrieval preference among candidate documents, we convert the scores into probability distributions using softmax. This allows us to determine whether adding the recovery component increases the relative probability assigned to the relevant document.

$$p_i^F = \mathrm{softmax}\left(\frac{s^0_{F,i}}{\tau}\right), \qquad p_i^U = \mathrm{softmax}\left(\frac{s^0_{F,i} + s^0_{P,i}}{\tau}\right). \tag{9}$$

The first distribution uses only the original full-space score, whereas the second additionally includes the Rank 1 recovery contribution. We activate the recovery signal only when this change increases the probability assigned to the relevant document, and define the selected target distribution as

$$g_i = \mathbb{1}\left[p^U_{i,0} > p^F_{i,0}\right], \qquad p_i^{\mathrm{sel}} = \mathrm{softmax}\left(\frac{s^0_{F,i} + g_i s^0_{P,i}}{\tau}\right). \tag{10}$$

Thus, $p_i^{\mathrm{sel}} = p_i^U$ when the recovery direction increases the relative probability of the relevant document, and $p_i^{\mathrm{sel}} = p_i^F$ otherwise. The trainable encoder produces normalized embeddings q_i^θ and $d^\theta_{i,j}$. Its retrieval scores are converted into a probability distribution, which is matched to the selected target distribution using KL divergence (Kullback & Leibler, 1951):

$$s^\theta_{ij} = q_i^{\theta\top} d^\theta_{i,j}, \qquad p_i^\theta = \mathrm{softmax}\left(\frac{s_i^\theta}{\tau}\right), \qquad \mathcal{L} = \frac{1}{B}\sum_{i=1}^{B} D_{\mathrm{KL}}\left(p_i^{\mathrm{sel}} \,\|\, p_i^\theta\right). \tag{11}$$

Here, B denotes the number of queries in a training batch. The target distribution is constructed from the frozen pretrained embeddings, while only the encoder parameters θ are updated during training.

5 Experiments

Implementation Details. We fine tune each pretrained encoder using a shared encoder for queries and documents with L2 normalized embeddings. Training uses an effective batch size of 128 queries for three epochs with AdamW, a learning rate of 10^{-5}, weight decay of 0.01, and gradient clipping with a maximum norm of 1.0. The learning rate follows a linear schedule with approximately 10% warmup followed by linear decay, and training is performed with BF16 mixed precision. Query and document inputs are truncated to 192 tokens during optimization and target distribution construction, while subspace construction and retrieval evaluation use a maximum length of 512 tokens. We use the 192 token limit to reduce training cost, while fully covering approximately 71–72% of corpus documents and remaining close to the average document length. This limit is therefore treated as a practical training setting rather than an empirically optimized hyperparameter. The temperature τ is set to a fixed value for the pretrained model; the motivation, selected value, and additional analysis are provided in the Appendix.

Candidate Construction. Each training instance contains one relevant document and seven explicit negative documents, together with in batch candidates. For the main training recipe, explicit negatives are selected from pretrained dense retrieval results whose scores are lower than 0.95 times the score of the selected positive document. If seven negatives cannot be obtained after extending retrieval to deeper ranks, the remaining candidates are randomly sampled after excluding known relevant documents and previously selected negatives. Our method modifies the supervision objective rather than the negative sampling strategy. We additionally combine it with representative negative samplers to examine whether its effect is maintained across different candidate distributions. For each sampler, standard fine tuning and our method use the same candidate construction protocol.

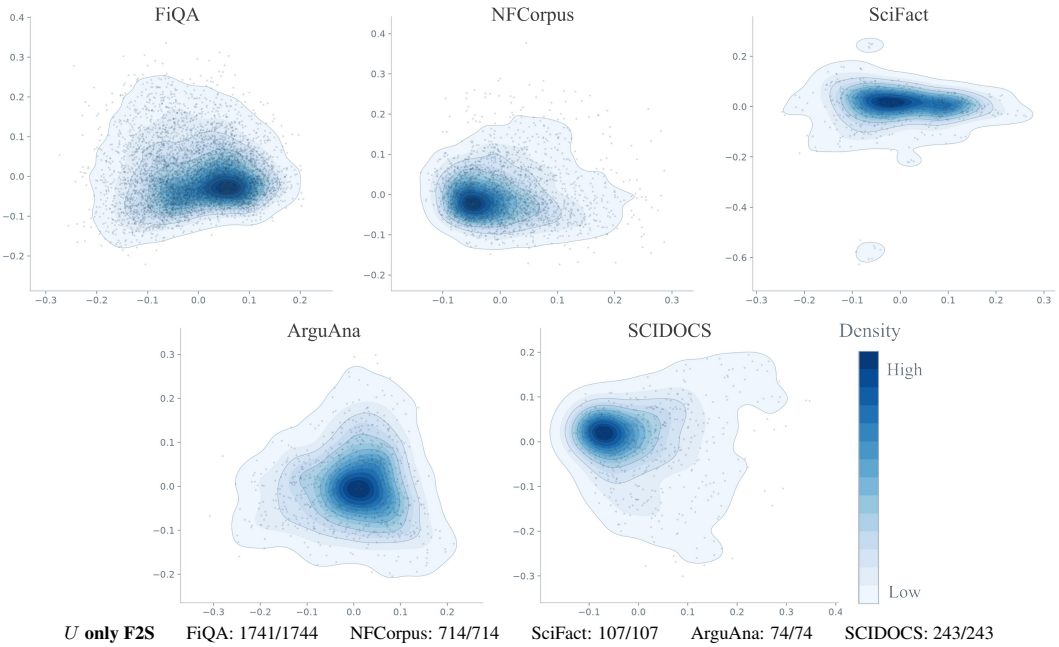

Figure 2: PCA visualization of query specific Rank 1 recovery projectors. Values show U only recovery among full space failures.

Experimental Setup. We evaluate the proposed method on six retrieval datasets from BEIR (Thakur et al., 2021): FiQA (Maia et al., 2018), NFCorpus (Boteva et al., 2016), SciFact (Wadden et al., 2020), ArguAna (Wachsmuth et al., 2018), SCIDOCS (Cohan et al., 2020). Experiments use Qwen3-Embedding-0.6B (Zhang et al., 2025) as pretrained embedding model. Standard fine tuning serves as the main training baseline and optimizes the InfoNCE contrastive objective (van den Oord et al., 2018; Karpukhin et al., 2020) using one relevant document and sampled negative candidates. We additionally evaluate standard fine tuning and our method with representative negative samplers, using the same candidate construction protocol for both methods under each sampler. Checkpoints are selected according to validation nDCG@10, with Recall@10 used to break ties. We report nDCG@10, Hit@10, and Recall@10 on the test set.

5.1 Recovery Projector Distribution and Recoverability

Figure 2 visualizes how the query specific Rank 1 recovery directions are distributed across the training queries. Since u_i and $-u_i$ represent the same Rank 1 subspace, we represent each recovery direction using its corresponding projector and apply PCA in the resulting projector space.

Each observation corresponds to one training query, and the height and color of each histogram bin represent the number of query specific projectors in that region. The visualization provides a low dimensional view of the recovery subspace distribution in the high dimensional projector space. Although each recovery direction is constructed separately for each query, many projectors appear in similar regions of the visualization, suggesting shared distributional patterns across training queries.

We further examine whether each constructed direction can recover queries that originally fail under full space retrieval. All failures are recovered on NFCorpus, SciFact, ArguAna, and SCIDOCS, while 1,741 of 1,744 failures are recovered on FiQA. This shows that the constructed Rank 1 directions provide a beneficial recovery signal for almost all failed queries. This high recoverability is closely related to the gate activation results presented in the following section.

Table 4: Main retrieval results across five datasets. Fine-tuned results are reported as mean ± standard deviation across 5 seeds.

Dataset	Method	nDCG@10	Hit@10	Recall@10
FiQA	Pretrained	46.935	75.772	55.005
	Standard FT	50.979 ± 0.115	79.064 ± 0.496	58.187 ± 0.298
	Ours	**51.475 ± 0.250**	**79.707 ± 0.527**	**59.424 ± 0.451**
NFCorpus	Pretrained	35.756	72.446	16.973
	Standard FT	40.107 ± 0.229	75.232 ± 0.819	**19.848 ± 0.205**
	Ours	**40.640 ± 0.268**	**76.367 ± 0.179**	19.820 ± 0.066
SciFact	Pretrained	69.894	83.333	82.389
	Standard FT	79.233 ± 0.649	91.667 ± 0.333	90.506 ± 0.347
	Ours	**80.338 ± 0.210**	**92.556 ± 0.192**	**91.944 ± 0.192**
ArguAna†	Pretrained	69.871	95.714	95.714
	Standard FT	**78.664 ± 0.556**	97.302 ± 0.275	97.302 ± 0.275
	Ours	76.509 ± 0.972	**98.095 ± 0.000**	**98.095 ± 0.000**
SCIDOCS†	Pretrained	23.455	64.667	24.167
	Standard FT	24.043 ± 0.162	**67.556 ± 0.770**	25.311 ± 0.336
	Ours	**25.012 ± 0.490**	65.778 ± 0.385	**26.163 ± 0.613**
Avg. Δ		**+0.181**	**+0.336**	**+0.858**

Green/red cells indicate gains/losses over Standard FT. Avg. Δ is the macro average of Ours − Standard FT. † Custom held out split.

5.2 Main Results

Table 4 compares the proposed method with the pretrained model and standard fine tuning across five retrieval datasets. Our method improves nDCG@10 on four of the five datasets and achieves better results on 12 of the 15 dataset and metric combinations. The macro average gains over standard fine tuning are $+0.181$ in nDCG@10, $+0.336$ in Hit@10, and $+0.858$ in Recall@10. The improvements are not uniform across all metrics, with lower nDCG@10 on ArguAna, slightly lower Recall@10 on NFCorpus, and lower Hit@10 on SCIDOCS.

We further examine these results through an ablation study. As shown in Section 5.1, the constructed Rank 1 directions recover almost all queries that fail under full space retrieval. Based on this result, we first examine how frequently the gate is actually activated during training. We then set $g = 0$ to determine whether the recovery signal itself contributes to learning. For datasets where the gate activation rate is below 100%, we additionally evaluate $g = 1$ to examine the effect of always using the recovery signal. When the activation rate is already 100%, the $g = 1$ setting is equivalent to the original setting and is therefore omitted.

Table 5: Ablation of the recovery contribution and gate activation.

Dataset	Method	Gate Rate	nDCG@10	Hit@10	Recall@10
FiQA	Ours	98.28%	51.475 ± 0.250	79.707 ± 0.527	59.424 ± 0.451
	$g = 0$	0%	46.863 ± 0.183	75.617 ± 0.408	54.400 ± 0.357
	$g = 1$	100%	51.432 ± 0.383	78.652 ± 0.624	58.658 ± 0.883
NFCorpus	Ours	100%	40.640 ± 0.268	76.367 ± 0.179	19.820 ± 0.066
	$g = 0$	0%	36.096 ± 0.116	72.549 ± 0.179	17.131 ± 0.071
SciFact	Ours	100%	80.338 ± 0.210	92.556 ± 0.192	91.944 ± 0.192
	$g = 0$	0%	69.967 ± 0.231	83.222 ± 0.385	82.233 ± 0.367
ArguAna†	Ours	99.74%	76.509 ± 0.972	98.095 ± 0.000	98.095 ± 0.000
	$g = 0$	0%	70.532 ± 0.272	95.714 ± 0.000	95.714 ± 0.000
	$g = 1$	100%	76.340 ± 1.151	98.095 ± 0.000	98.095 ± 0.000
SCIDOCS†	Ours	100%	25.012 ± 0.490	65.778 ± 0.385	26.163 ± 0.613
	$g = 0$	0%	23.270 ± 0.237	65.556 ± 0.770	24.311 ± 0.250

Dark green indicates the best result within each dataset and metric. $g = 0$ removes the recovery contribution, while $g = 1$ applies it to all queries. † Custom held out split.

Table 5 shows that removing the recovery signal with $g = 0$ decreases all three retrieval metrics across all five datasets. This consistent degradation indicates that the Rank 1 recovery signal is a main source of the improvement provided by the proposed supervision.

Consistent with the high recoverability observed in Section 5.1, the gate is activated for most training queries, with activation rates ranging from 98.28% to 100%. For FiQA and ArguAna, where the gate is not fully activated, forcing $g = 1$ produces similar results but slightly decreases performance. These results indicate that most constructed recovery directions provide useful supervision, while the gate mainly prevents the small fraction of unsuitable recovery signals from being added to the target distribution.

5.3 Compatibility with Negative Mining Strategies

As discussed in the candidate construction, the proposed supervision is defined separately from the negative mining procedure and can be applied to different candidate sets. We therefore examine whether its effect is maintained when combined with different negative mining strategies. For this analysis, we focus on NFCorpus, SciFact, and SCIDOCS, where at least one negative mining setting achieves higher nDCG@10 than Standard FT.

We compare Standard FT with SimANS (Zhou et al., 2022) and ANCE (Xiong et al., 2021), and apply the proposed supervision to each setting. This allows us to examine whether the recovery signal provides an additional benefit when the negative mining strategy itself is changed.

Table 6: Results with different negative sampling strategies.

Dataset	Method	nDCG@10		Hit@10		Recall@10	
		Base	+ Ours	Base	+ Ours	Base	+ Ours
NFCorpus	Standard FT	40.107 ± 0.229	40.640 ± 0.268	75.232 ± 0.819	76.367 ± 0.179	19.848 ± 0.205	19.820 ± 0.066
	SimANS	40.121 ± 0.398	40.630 ± 0.155	76.161 ± 0.819	76.677 ± 0.995	19.902 ± 0.064	19.653 ± 0.114
	ANCE	40.117 ± 0.229	40.434 ± 0.123	75.542 ± 0.536	75.851 ± 0.536	19.720 ± 0.397	19.621 ± 0.160
SciFact	Standard FT	79.233 ± 0.649	80.338 ± 0.210	91.667 ± 0.333	92.556 ± 0.192	90.506 ± 0.347	91.944 ± 0.192
	SimANS	76.548 ± 2.276	80.143 ± 0.099	90.444 ± 2.411	93.000 ± 0.000	89.926 ± 2.611	92.400 ± 0.084
	ANCE	79.950 ± 0.421	80.560 ± 0.320	91.889 ± 0.385	92.444 ± 0.385	90.706 ± 0.482	91.833 ± 0.536
SCIDOCS†	Standard FT	24.043 ± 0.162	25.012 ± 0.490	67.556 ± 0.770	65.778 ± 0.385	25.311 ± 0.336	26.163 ± 0.613
	SimANS	24.688 ± 0.308	24.892 ± 0.456	69.111 ± 0.385	66.444 ± 0.385	25.867 ± 0.283	25.344 ± 0.684
	ANCE	24.284 ± 0.092	24.953 ± 0.322	68.000 ± 0.667	68.222 ± 0.385	25.533 ± 0.154	26.078 ± 0.212

Dark green indicates the best result within each dataset and metric.
Light green/red indicates gains/losses after adding the proposed supervision. † Custom held out split.

Table 6 reports the results under different negative mining strategies. Across the three datasets, adding our supervision improves nDCG@10 for all Standard FT, SimANS, and ANCE settings. The improvement is especially clear on SciFact, where SimANS increases from 76.548 to 80.143 and ANCE increases from 79.950 to 80.560 after adding our supervision. On NFCorpus and SCIDOCS, the gains in nDCG@10 are smaller but remain consistent across the tested settings.

The behavior of Hit@10 and Recall@10 is more mixed. SimANS with our supervision improves all three metrics on SciFact, while decreases are observed in Recall@10 on NFCorpus and in Hit@10 and Recall@10 for SimANS on SCIDOCS. Overall, these results show that the recovery signal can provide additional improvements beyond the standard fine tuning setting under different negative mining strategies.

6 Discussion and Conclusion

Our analysis shows that retrieval failure in the full embedding space does not necessarily indicate that useful retrieval structure is absent from the frozen representation. For many failed queries, relevant documents can still be recovered by restricting retrieval to low dimensional subspaces. We further find that components outside these recovery subspaces can interfere with the recovery signal, while smaller recovery subspaces can contain disproportionately large amounts of discriminative information relative to their dimensionality.

Based on these observations, we propose Selective Low Rank Supervision, which derives a query specific Rank 1 recovery direction and selectively incorporates its contribution into the training target distribution. The recovery direction is used only during training, while inference remains unchanged and uses standard full space similarity. Across five retrieval datasets, the proposed supervision generally improves over standard fine tuning, and ablation results show that removing the recovery contribution consistently reduces performance. Because the proposed supervision is defined independently of the contrastive learning objective and negative mining procedure, it can be combined with different contrastive learning and candidate construction strategies.

References

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

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

João Coelho, Bruno Martins, Joao Magalhaes, Jamie Callan, and Chenyan Xiong. Dwell in the beginning: How language models embed long documents for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 370–377. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-short.35. URL https://aclanthology.org/2024.acl-short.35/.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.207. URL https://aclanthology.org/2020.acl-main.207/.

Chao-Wei Huang and Yun-Nung Chen. PairDistill: Pairwise relevance distillation for dense retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18225–18237. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1013. URL https://aclanthology.org/2024.emnlp-main.1013/.

Junfeng Kang, Rui Li, Qi Liu, Zhenya Huang, Zheng Zhang, Yanjiang Chen, Linbo Zhu, and Yu Su. Distribution-driven dense retrieval: Modeling many-to-one query-document relationship. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11933–11941, 2025. doi: 10.1609/aaai.v39i11.33299. URL https://doi.org/10.1609/aaai.v39i11.33299.

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 (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Jaeyoung Kim, Dohyeon Lee, and Seung-won Hwang. HIL: Hybrid isotropy learning for zero-shot performance in dense retrieval. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 7892–7903. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.437. URL https://aclanthology.org/2024.naacl-long.437/.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

Yifan Liu, Qianfeng Wen, Mark Zhao, Jiazhou Liang, and Scott Sanner. MA-DPR: Manifold-aware distance metrics for dense passage retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31085–31103. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1582. URL https://aclanthology.org/2025.emnlp-main.1582/.

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

Zhijie Nie, Richong Zhang, and Zhanyu Wu. A text is worth several tokens: Text embedding from LLMs secretly aligns well with the key tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7683–7694. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.379. URL https://aclanthology.org/2025.acl-long.379/.

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

Minh-Phuc Truong, Hai An Vu, Tu Vu, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen, and Trung Le. EMO: Embedding model distillation via intra-model relation and optimal transport alignments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7594–7606. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.385. URL https://aclanthology.org/2025.emnlp-main.385/.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Robin Vujanic and Thomas Rückstieß. LEAF: Knowledge distillation of text embedding models with teacher-aligned representations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43362–43383. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2008. URL https://aclanthology.org/2026.acl-long.2008/.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 241–251. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1023. URL https://aclanthology.org/P18-1023/.

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

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10007818.

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://iclr.cc/virtual/2021/poster/2673.

Liyan Xu, Zhenlin Su, Mo Yu, Jiangnan Li, Fandong Meng, and Jie Zhou. Dense retrievers can fail on simple queries: Revealing the granularity dilemma of embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19295–19305. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1051. URL https://aclanthology.org/2025.findings-emnlp.1051/.

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.

Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-rong Wen, and Nan Duan. SimANS: Simple ambiguous negatives sampling for dense text retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 548–559. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-industry.56. URL https://aclanthology.org/2022.emnlp-industry.56/.

Yucheng Zhou, Tao Shen, Xiubo Geng, Chongyang Tao, Jianbing Shen, Guodong Long, Can Xu, and Daxin Jiang. Fine-grained distillation for long document retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19732–19740, 2024. doi: 10.1609/aaai.v38i17.29947. URL https://doi.org/10.1609/aaai.v38i17.29947.

AI USE STATEMENT

Generative AI tools were used to assist with experiment orchestration, result aggregation, and language editing during manuscript preparation. The research questions, methodological design, experimental setup, interpretation of the results, and conclusions were determined by the authors. All reported results, equations, tables, figures, and citations were reviewed against the corresponding experiment artifacts and source materials by the authors.

ETHICS STATEMENT

This study uses existing benchmark retrieval datasets and does not collect new human-subject data. Relevance judgments in retrieval datasets may be incomplete, and documents treated as nonrelevant during candidate construction are therefore not interpreted as definitively irrelevant. Dataset splits and task-specific adaptations used in our supervised experiments are explicitly described to distinguish them from official benchmark settings. The study focuses on representation and retrieval behavior and does not evaluate deployment-specific fairness, privacy, or downstream societal effects.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the recovery subspace construction, Rank 1 recovery direction, selective target distribution, dataset partitions, candidate construction, negative mining strategies, and training configuration. The Appendix provides additional rank-wise recovery analyses, residual suppression results, subspace geometry analyses, additional ablations, and negative mining experiments.

A RANK-WISE ORACLE RECOVERY ACROSS DATASETS

We evaluate frozen Qwen3-Embedding-0.6B representations ($D = 1024$) using 1,024 Gaussian-QR candidate subspaces for each rank $r \in \{1, 2, 4, 8, 16, 32\}$. The rank-specific candidate pool is shared across datasets. Scores are computed as $q^\top UU^\top d$, without post-projection normalization. Recovery is evaluated over the full corpus. The archived analysis model metadata specifies a maximum input length of 8,192 tokens; these caches are distinct from the 512-token caches used for the later constructed-U training analysis. These results are label-conditioned oracle unions over candidates, rather than performance of a learned subspace selector. For each rank, Unique F2S counts baseline-failed queries recovered by at least one candidate of that rank. Total success is the union of baseline successes and that rank's recoveries. The All ranks row unions recoveries across all six ranks without double counting. The analysis query sets differ from the custom supervised held-out splits used in the fine-tuning experiments.

Table 7: Rank-wise oracle recovery on FiQA using frozen Qwen3-Embedding-0.6B.

Rank	Unique F2S@10	Total success@10	Success (%)
Baseline	–	492/648	75.93
1	49	541/648	83.49
2	84	576/648	88.89
4	111	603/648	93.06
8	136	628/648	96.91
16	147	639/648	98.61
32	150	642/648	99.07
All ranks	155	647/648	99.85

B ORTHOGONAL DECOMPOSITION OF RETRIEVAL SCORES

Let $U \in \mathbb{R}^{D\times r}$ have orthonormal columns, $U^\top U = I_r$, and let $P = UU^\top$ and $R = I_D - P$.

Table 8: Rank-wise oracle recovery on NFCorpus using frozen Qwen3-Embedding-0.6B.

Rank	Unique F2S@10	Total success@10	Success (%)
Baseline	–	235/323	72.76
1	87	322/323	99.69
2	87	322/323	99.69
4	87	322/323	99.69
8	87	322/323	99.69
16	86	321/323	99.38
32	86	321/323	99.38
All ranks	88	323/323	100.00

Table 9: Rank-wise oracle recovery on SciFact using frozen Qwen3-Embedding-0.6B.

Rank	Unique F2S@10	Total success@10	Success (%)
Baseline	–	251/300	83.67
1	48	299/300	99.67
2	44	295/300	98.33
4	49	300/300	100.00
8	49	300/300	100.00
16	49	300/300	100.00
32	49	300/300	100.00
All ranks	49	300/300	100.00

Projection properties. Direct calculation gives

$$P^\top = P, \quad P^2 = U(U^\top U)U^\top = P, \quad R^\top = R, \quad R^2 = R, \quad PR = RP = 0.$$

Thus P and R are complementary orthogonal projectors. The identity $I_D = P + R$ is algebraic; orthogonality additionally requires the projector properties above. Every $z \in \mathbb{R}^D$ satisfies

$$z = Pz + Rz, \qquad (Pz)^\top Rz = 0, \qquad \|z\|_2^2 = \|Pz\|_2^2 + \|Rz\|_2^2.$$

Consequently, for any query q and document d,

$$q^\top d = (Pq + Rq)^\top (Pd + Rd) \tag{12}$$

$$= (Pq)^\top (Pd) + (Rq)^\top (Rd) \tag{13}$$

$$= q^\top Pd + q^\top Rd. \tag{14}$$

The cross terms vanish because $PR = RP = 0$. In particular, for $\Delta = d^+ - d^-$,

$$m_F = q^\top \Delta = m_P + m_R, \qquad m_P = q^\top P\Delta, \quad m_R = q^\top R\Delta.$$

If $m_P > 0$ but $m_F \leq 0$, then $m_R \leq -m_P$: the residual cancels the selected component for this fixed document pair. This pairwise statement alone does not determine full-corpus top-k success.

Residual suppression. For $0 \leq \alpha \leq 1$,

$$s_\alpha = q^\top Pd + (1-\alpha) q^\top Rd = (1-\alpha) q^\top d + \alpha q^\top Pd.$$

Hence $s_0 = q^\top d$ and $s_1 = q^\top Pd$. Equivalently, $s_\alpha = (T_\alpha q)^\top (T_\alpha d)$ with $T_\alpha = P + \sqrt{1-\alpha}\, R$. Applying $P + (1-\alpha)R$ to both embeddings instead would square the residual coefficient. These identities concern raw dot products; separately normalizing the projected embeddings changes the scores and does not preserve this additive decomposition. Neither the identity nor linear interpolation guarantees monotonic top-k recovery as α varies.

C Dataset-wise Effects of Residual Suppression

We reuse the frozen Qwen embeddings and the 1,024 candidates per rank described in Appendix A. For each query and rank, we select the candidate with the smallest best-relevant-document rank,

Table 10: Rank-wise oracle recovery on SCIDOCS using frozen Qwen3-Embedding-0.6B.

Rank	Unique F2S@10	Total success@10	Success (%)
Baseline	–	661/1000	66.10
1	277	938/1000	93.80
2	307	968/1000	96.80
4	324	985/1000	98.50
8	333	994/1000	99.40
16	337	998/1000	99.80
32	336	997/1000	99.70
All ranks	339	1000/1000	100.00

breaking ties by the largest margin between the frozen baseline positive and negative, then the smallest candidate ID. This label-conditioned selection is performed once and held fixed throughout suppression. For $\alpha \in \{0.3, 0.5, 0.9\}$, we evaluate $s_\alpha = (1-\alpha)s_F + \alpha s_P$ over the full corpus. The $\alpha = 0$ and $\alpha = 1$ endpoints are checked against the archived full and selected-projection success outcomes.

F2S queries denotes the initially failed queries recovered by the selected pure projection. The three suppression columns count successes within this fixed cohort. Max. S2F is the maximum number of initially successful queries that fail at any one of the three listed coefficients, evaluated over all baseline-successful queries; it is not the number of unique queries that fail anywhere on the path. The CSV also records S2F restricted to queries successful at both endpoints. This explicitly specified Qwen protocol is a new extension, not a numerical reproduction of the E5 experiment underlying the manuscript table.

Table 11: Residual suppression on FiQA. Max. S2F uses all 492 baseline-successful queries.

Rank	F2S queries	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.9$	Max. S2F
1	49	1	1	9	1
2	84	2	4	25	4
4	111	4	10	39	1
8	136	4	9	63	1
16	147	6	12	85	0
32	150	9	18	100	0

Table 12: Residual suppression on NFCorpus. Max. S2F uses all 235 baseline-successful queries.

Rank	F2S queries	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.9$	Max. S2F
1	87	0	1	13	0
2	87	0	0	19	0
4	87	0	3	26	0
8	87	1	5	32	0
16	86	2	6	42	0
32	86	4	11	48	0

The decomposition does not imply that residual suppression is harmless for every query. Nonzero S2F counts occur in these Qwen experiments and should be reported alongside recovery gains. Subspaces are selected with relevance labels, so these interventions are oracle diagnostics.

D Dataset-wise Subspace Overlap and Concentration

For each baseline-failed query recovered by at least two distinct candidates of rank r, we average $\|U_a^\top U_b\|_F^2/r$ over all unordered successful-candidate pairs. We then give equal weight to eligible queries. The table reports the eligible query count $N_{\geq 2}$, which can be smaller than Unique F2S. The random control is the exact mean over all distinct pairs in the archived rank-specific pool; this

Table 13: Residual suppression on SciFact. Max. S2F uses all 251 baseline-successful queries.

Rank	F2S queries	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.9$	Max. S2F
1	48	0	1	16	0
2	44	0	2	18	0
4	49	0	3	23	0
8	49	2	4	27	0
16	49	2	6	31	0
32	49	3	9	33	0

Table 14: Residual suppression on SCIDOCS. Max. S2F uses all 661 baseline-successful queries.

Rank	F2S queries	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.9$	Max. S2F
1	277	1	2	22	7
2	307	3	7	45	7
4	324	4	10	69	3
8	333	7	9	109	10
16	337	10	24	139	2
32	336	12	31	192	0

equals the expected pair-average for a uniformly sampled subset of any fixed size at least two. It is therefore a matched-size null expectation without Monte Carlo subset noise. Ratios are descriptive, conditional on the shared pool; no significance claim is made.

For discriminative concentration, we reuse the archived occupancy of one oracle-selected candidate per recovered query: smallest relevant-document rank, breaking ties by the first candidate ID. This selection rule differs from the margin tie-break used in the suppression analysis. The document difference is between the frozen baseline best relevant and hardest nonrelevant document. We average $\|P(d^+ - d^-)\|^2/\|d^+ - d^-\|^2$ equally over queries with a nonzero difference. Expected concentration is r/D with $D = 1024$; enrichment is the ratio of the displayed unrounded means. Overlap and concentration thus have different eligibility and aggregation rules.

Table 15: Overlap and discriminative concentration on FiQA. Disc. and expected values are percentages; ratios and enrichment are dimensionless.

Rank	$N_{\geq 2}$	Success-U	Random	Ratio	Disc.	Expected	Enrich.
1	15	0.000484	0.000977	0.495	0.253	0.098	2.586
2	35	0.002053	0.001952	1.052	0.346	0.195	1.771
4	71	0.003888	0.003906	0.995	0.679	0.391	1.738
8	109	0.007877	0.007815	1.008	0.978	0.781	1.252
16	139	0.015678	0.015624	1.003	1.692	1.562	1.083
32	145	0.031273	0.031251	1.001	3.223	3.125	1.031

Dimensional expectation. For an isotropic random rank-r orthogonal projector P, rotational invariance implies $\mathbb{E}[P] = aI_D$. Taking traces gives $a = r/D$. Hence, for fixed $v \neq 0$, $\mathbb{E}[\|Pv\|^2/\|v\|^2] = r/D$. Similarly, for two independent isotropic rank-r subspaces, $\mathbb{E}[\mathrm{tr}(P_a P_b)/r] = r/D$. The empirical shared-pool control above need not equal this population expectation exactly.

Rank-one discriminative enrichment exceeds one in all five datasets. In contrast, successful-subspace overlap does not uniformly exceed the random control: the rank-one ratios are below one for FiQA, NFCorpus, and SciFact. Consequently, increased overlap is dataset dependent and should not be asserted as a universal Qwen property.

Table 16: Overlap and discriminative concentration on NFCorpus. Disc. and expected values are percentages; ratios and enrichment are dimensionless.

Rank	$N_{\geq 2}$	Success-U	Random	Ratio	Disc.	Expected	Enrich.
1	84	0.000929	0.000977	0.951	0.269	0.098	2.751
2	85	0.001923	0.001952	0.985	0.352	0.195	1.800
4	85	0.003945	0.003906	1.010	0.484	0.391	1.239
8	85	0.007833	0.007815	1.002	0.903	0.781	1.156
16	84	0.015644	0.015624	1.001	1.667	1.562	1.067
32	83	0.031204	0.031251	0.999	3.029	3.125	0.969

Table 17: Overlap and discriminative concentration on SciFact. Disc. and expected values are percentages; ratios and enrichment are dimensionless.

Rank	$N_{\geq 2}$	Success-U	Random	Ratio	Disc.	Expected	Enrich.
1	34	0.000887	0.000977	0.908	0.446	0.098	4.564
2	37	0.002069	0.001952	1.060	0.545	0.195	2.792
4	48	0.004051	0.003906	1.037	0.634	0.391	1.623
8	49	0.007873	0.007815	1.008	0.999	0.781	1.279
16	47	0.015666	0.015624	1.003	1.810	1.562	1.159
32	48	0.031208	0.031251	0.999	3.028	3.125	0.969

E Supporting Results for Recovery-Based Supervision

We distinguish properties of a labeled construction pair from properties of retrieval over an entire candidate set. Let q, d^+, d^- be unit vectors, $\Delta = d^+ - d^- \neq 0$, $\delta = \Delta/\|\Delta\|_2$, and $c = q^\top \delta$. Assume $c > -1$ so that $u = (q+\delta)/\|q+\delta\|_2$ is defined.

Optimal unit separation direction. For every unit vector v, Cauchy–Schwarz gives

$$v^\top (d^+ - d^-) \leq \|d^+ - d^-\|_2.$$

Equality holds for $v = \delta$. Thus δ maximizes the signed document-pair separation under a unit-norm constraint. This is not yet a query-conditioned retrieval statement.

Query-conditioned projected separation. Define $A = (q\delta^\top + \delta q^\top)/2$. For a unit vector v,

$$(q^\top v)(v^\top \delta) = v^\top A v.$$

For $-1 < c < 1$, the normalized vectors proportional to $q+\delta$ and $q-\delta$ are orthogonal eigenvectors with eigenvalues $(1+c)/2$ and $(c-1)/2$, respectively. All directions orthogonal to their span have eigenvalue zero. For $c = 1$, $A = qq^\top$ and $u = q$. Therefore, by the Rayleigh quotient,

$$u \in \arg\max_{\|v\|_2=1} (q^\top v)(v^\top \delta), \qquad \max_{\|v\|_2=1} v^\top A v = \frac{1+c}{2}.$$

For the construction pair specifically,

$$q^\top u u^\top (d^+ - d^-) = \|\Delta\|_2 \frac{1+c}{2}, \qquad q^\top (d^+ - d^-) = \|\Delta\|_2 c.$$

Thus the projected pair margin exceeds or equals the full-space pair margin by $\|\Delta\|_2(1-c)/2$. Adding the projected score increases this pair margin by $\|\Delta\|_2(1+c)/2$. These results do not guarantee improvement against every other document. They assume that the same embeddings define u and the scores; they do not automatically apply to the 192-token scoring embeddings when u was constructed from 512-token embeddings.

Exact meaning of the gate. For a fixed candidate set, let $a_j = s_{F,j}/\tau$, $b_j = s_{P,j}/\tau$, and index the selected positive by 0. For $\tau > 0$,

$$\frac{p_0^U}{p_0^F} = \frac{\exp(b_0)}{\sum_j p_j^F \exp(b_j)}.$$

Table 18: Overlap and discriminative concentration on SCIDOCS. Disc. and expected values are percentages; ratios and enrichment are dimensionless.

Rank	$N_{\geq 2}$	Success-U	Random	Ratio	Disc.	Expected	Enrich.
1	182	0.001006	0.000977	1.029	0.224	0.098	2.292
2	239	0.001987	0.001952	1.018	0.319	0.195	1.632
4	279	0.003904	0.003906	0.999	0.497	0.391	1.273
8	317	0.007858	0.007815	1.006	0.832	0.781	1.064
16	327	0.015587	0.015624	0.998	1.659	1.562	1.061
32	327	0.031270	0.031251	1.001	3.161	3.125	1.011

Consequently, the gate activates exactly when $\exp(b_0) > \sum_j p_j^F \exp(b_j)$. The gate compares normalized positive probabilities rather than the positive logit alone. With the stated strict comparison,

$$p_0^{\text{sel}} = \max\{p_0^F, p_0^U\}.$$

This is a candidate-set guarantee for the frozen target. It does not imply improved positive rank, nDCG, or generalization by the trained encoder. At equality the full-space distribution is selected. Stable log-probability comparison avoids softmax saturation changing the intended mathematical test; this is a numerical consideration, not a new gate threshold.

KL supervision and the encoder gradient. For a fixed target $t_j = p_j^{\text{sel}}$ and student logits $z_j = s_j^\theta/\tau$,

$$\mathcal{L} = \sum_j t_j(\log t_j - \log \operatorname{softmax}(z)_j), \qquad \frac{\partial \mathcal{L}}{\partial s_j^\theta} = \frac{p_j^\theta - t_j}{\tau}.$$

The batch-averaged gradient includes an additional factor $1/B$. Gradients do not flow through the frozen target. If $g = 0$ and student and teacher distributions agree exactly, this loss gives zero gradient for that instance. After shared encoder updates change the student distribution, the same target can supply a restorative gradient. Dropout, differing inputs and other optimizer effects must be distinguished from this exact statement. Inference uses only the trained encoder and full-space dot products; the construction is not an additional inference-time component.

F Numerical Data Underlying Figure 2

Figure 2 uses all cached training queries and the fixed analytic directions used for F. Each point corresponds to a query-specific projector, not a random candidate subspace. Full-corpus retrieval uses raw projected dot products, excludes a nonrelevant same-ID document, and breaks score ties by corpus index. These are label-informed training-set oracle outcomes, not held-out encoder performance.

Table 19: Exact recovery counts underlying Figure 2. Recovery percentage is F2S divided by the number of baseline failures.

Dataset	Queries	Failures	F2S	Recovery (%)	S2F	F2F
FiQA	5500	1744	1741	99.828	2	3
NFCorpus	2582	714	714	100.000	0	0
SciFact	726	107	107	100.000	0	0
ArguAna	981	74	74	100.000	2	0
SCIDOCS	700	243	243	100.000	0	0

Projector PCA. Let $P_i = u_i u_i^\top$. The full-vectorized projector inner product is $\langle P_i, P_j \rangle_F = (u_i^\top u_j)^2$. We perform centered kernel PCA using this Gram matrix, equivalent to ordinary PCA on full flattened projectors, without materializing D^2 entries per query. This avoids the sign ambiguity of u_i. PCA is fitted independently for each dataset; axes are not a common coordinate system across datasets. The plot uses 14×14 histogram bins; bar height and color encode counts, not a recovery probability or KDE.

Table 20: Explained variance of the two plotted projector principal components.

Dataset	PC1 (%)	PC2 (%)	Total (%)
FiQA	0.688	0.571	1.259
NFCorpus	0.513	0.425	0.939
SciFact	0.802	0.717	1.519
ArguAna	0.694	0.628	1.322
SCIDOCS	0.854	0.638	1.492

Only a small fraction of projector variance is represented in two dimensions. Concentration in the figure alone does not establish high-dimensional clustering or collapse. Complete per-query recovery records and histogram bin counts accompany the summary tables.

G Why the Constructed Projector Has Rank One

Assume $d^+ \neq d^-$ and $q + \delta \neq 0$, where $\delta = (d^+ - d^-)/\|d^+ - d^-\|_2$. Then $u = (q + \delta)/\|q + \delta\|_2$ is a single nonzero unit vector. Define $U = [u] \in \mathbb{R}^{D\times 1}$ and $P = UU^\top = uu^\top$. Since its sole column is nonzero, $\mathrm{rank}(U) = 1$. Moreover, for every x,

$$Px = u(u^\top x) \in \mathrm{span}\{u\}, \qquad Pu = u.$$

The image of P is therefore exactly $\mathrm{span}\{u\}$, so $\mathrm{rank}(P) = 1$. Equivalently, its eigenvalues are one along u and zero on $u^\perp$. Although constructing u uses three input embeddings, it forms one linear combination, not a three-column basis $[q, d^+, d^-]$. The span of those inputs can have dimension up to three; the selected subspace has dimension one by construction. This rank statement does not establish that rank one is the optimal rank for retrieval.

Degenerate inputs. If $d^+ = d^-$, δ is undefined; if $q = -\delta$, u is undefined. Such instances require an explicit exclusion or fallback rule before the normalized construction is used. An epsilon clamp alone does not turn the zero vector into a unit vector: using $u = 0$ gives a rank-zero projector. For $q = -\delta$ and $D \geq 2$, the maximum query-conditioned projected separation is zero and any unit vector perpendicular to q is a maximizer, but choosing one would be an additional convention. The theoretical rank-one claim applies to the stated nondegenerate domain. Finally, $(-u)(-u)^\top = uu^\top$; the orientation sign does not change the rank-one subspace or its retrieval scores.

H Additional Negative-Sampler Results

The following table completes the ANCE-style and SimANS comparisons for FiQA and ArguAna. Values are means and sample standard deviations over seeds 42, 43, and 44. ArguAna uses the custom supervised held-out split. Frozen teacher and original F subspaces are retained in the combined runs. ANCE-style uses exact global mining with the current encoder every half epoch. Its realized candidate identities can differ between the baseline and F runs because their encoders evolve differently. Baseline InfoNCE uses temperature 0.01 and F KL uses 0.05. These are comparisons of the implemented training recipes, not a temperature-matched loss-only ablation.

Table 21: Additional results with different negative mining strategies.

Dataset	Sampler	nDCG@10		Hit@10		Recall@10	
		Base	+F	Base	+F	Base	+F
FiQA	ANCE-style	48.842±0.398	51.484±0.168	77.006±0.154	78.292±0.729	55.925±0.311	58.318±0.521
	SimANS	49.603±0.491	51.287±0.153	78.344±0.388	78.447±0.761	57.083±0.303	58.498±0.680
ArguAna	ANCE-style	79.360±1.420	77.920±0.550	97.619±0.000	98.095±0.476	97.619±0.000	98.095±0.476
	SimANS	77.847±1.040	76.000±0.740	98.413±0.550	98.730±0.275	98.413±0.550	98.730±0.275

On FiQA, both sampler combinations improve all three metrics. On ArguAna, both combinations reduce nDCG while improving Hit and Recall. The added supervision therefore does not improve every dataset–sampler–metric combination. The main-table Standard FT rows are not duplicated here.

I LIMITS OF DIRECT SUBSPACE SUPERVISION

A useful oracle subspace does not by itself imply that directly optimizing projected scores is sufficient for full-space retrieval. We establish a lack of such a guarantee, rather than an impossibility of learning from subspaces, and report a completed pilot that measures both recovery and damage to previously successful queries.

I.1 DIRECT PROJECTED OBJECTIVES

For a fixed constructed direction u_i, let $P_i = u_i u_i^\top$ and $R_i = I - P_i$. The encoder is trainable, whereas P_i is held fixed. Write

$$s^\theta_{P,ij} = q_i^{\theta\top} P_i d_j^\theta, \qquad s^\theta_{R,ij} = q_i^{\theta\top} R_i d_j^\theta, \qquad s^\theta_{F,ij} = s^\theta_{P,ij} + s^\theta_{R,ij}.$$

The completed negative-free experiment uses

$$\mathcal{L}_{\text{direct}+} = \frac{1}{B}\sum_{i=1}^{B} \text{softplus}\left(-\frac{s^\theta_{P,i+}}{\tau}\right), \qquad \tau = 0.05, \tag{15}$$

where $\text{softplus}(x) = \log(1 + e^x)$. It updates the encoder through the projected positive score, with no teacher KL, explicit training negatives, or in-batch negative comparisons. A negative document is still used offline when constructing u_i; "negative-free" refers to the optimization objective, not to subspace construction. No normalization is applied after projection.

A distinct projected pairwise objective would be

$$\mathcal{L}_{\text{pair},P} = \text{softplus}(-m_P/\tau), \qquad m_P = q^\top P(d^+ - d^-).$$

This objective explicitly compares a positive and a negative. It is covered by the following argument, but is not the objective of the reported pilot.

I.2 PROJECTED SCORES DO NOT DETERMINE FULL-SPACE RANKINGS

Proposition (non-identifiability of full-space order). For a fixed proper rank-one projector, identical projected positive scores and projected positive–negative margins can correspond to opposite full-space rankings, even when all embeddings are unit normalized. Thus neither of the objectives above alone guarantees full-space Success@10.

Proof. Consider $P = \text{diag}(1, 0)$, $q = (0.6, 0.8)$, and two possible embedding states:

	d^+	d^-
A	$(0.6, 0.8)$	$(-0.6, -0.8)$
B	$(0.6, -0.8)$	$(-0.6, 0.8)$.

All vectors have unit norm. In both states,

$$s_P^+ = 0.36, \qquad s_P^- = -0.36, \qquad m_P = 0.72.$$

Consequently, both the positive-only and projected pairwise losses are identical between the states. Nevertheless,

$$(s_F^+, s_F^-)_A = (1, -1), \qquad (s_F^+, s_F^-)_B = (-0.28, 0.28).$$

The positive is preferred in state A and the negative in state B. The uncontrolled residual margin changes the sign of $m_F = m_P + m_R$.

To obtain a strict Success@10 counterexample with distinct documents, embed these vectors in $\mathbb{R}^3$ and choose ten distinct angles $0 < \theta_j < 0.1$. Set

$$d^-_{A,j} = (-0.6, -0.8\cos\theta_j, -0.8\sin\theta_j), \quad d^-_{B,j} = (-0.6, 0.8\cos\theta_j, 0.8\sin\theta_j).$$

Keep q, d_A^+, d_B^+ as above with a zero third coordinate and let $P = \text{diag}(1, 0, 0)$. Every projected negative score remains -0.36 in both states. In state A, all ten negatives score below the positive; in state B, all ten score above it, since $-0.36 + 0.64\cos\theta_j > -0.28$. The positive rank changes from 1 to 11 without changing either projected objective. □

This is a statement about what the loss constrains, not a claim that a particular optimizer must take this path. It also applies when P was constructed analytically at initialization and subsequently frozen: the construction-pair result in Appendix E is not a constraint on every future encoder state. Conversely, the example does not rule out successful direct subspace training under additional assumptions or objectives.

Further limitation of the positive-only objective. Equation (15) contains no competitor score, so it does not directly enforce discrimination even inside the selected subspace. A decrease in this loss can coexist with a larger increase in a nonrelevant document's score. Shared encoder parameters can affect these documents despite their absence from the batch.

Projected gradients are not a frozen residual encoder. At the level of independent embedding variables, the positive-only query gradient is proportional to Pd^+ and therefore lies in $\mathrm{span}(u)$. This does not mean that an encoder update leaves all residual embeddings unchanged. If $x = h/\|h\|$, the normalization Jacobian is $(I - xx^\top)/\|h\|$; backpropagation through it and the shared encoder need not preserve an ambient direction. Both normalization and parameter sharing distinguish a projected embedding gradient from a constraint on future full-space representations.

Component contrast is also a different objective. A negative-free comparison of two components of the same positive pair, such as

$$\ell_{\mathrm{comp}} = \mathrm{softplus}\left(\frac{\gamma - s_P^+ + s_R^+}{\tau}\right),$$

should not be confused with document-pair ranking. With $a = (\gamma - s_P^+ + s_R^+)/\tau$ and the sigmoid σ,

$$\frac{\partial \ell_{\mathrm{comp}}}{\partial s_P^+} = -\frac{\sigma(a)}{\tau}, \qquad \frac{\partial \ell_{\mathrm{comp}}}{\partial s_R^+} = \frac{\sigma(a)}{\tau}.$$

At the score level, reducing the residual positive contribution decreases this loss while also decreasing $s_F^+ = s_P^+ + s_R^+$. This illustrates a potential objective mismatch; it is not the loss used for the numerical result below, and no causal attribution to this mechanism is made.

I.3 Pilot: recovery and damage must both be counted

We reuse the completed Qwen3-Embedding-0.6B experiment on SciFact, seed 42. The direct arm uses the same initial checkpoint, selected training positives, and fixed analytic rank-one directions as F. It trains for 18 updates over three epochs with query batch size 128 and learning rate 10^{-5}. Training inputs are capped at 192 tokens; subspace construction and retrieval evaluation use 512 tokens. All arms select a checkpoint by validation nDCG@10 and use ordinary full-space retrieval on the 300-query official test split. The selected step is 6 for the direct arm and 18 for F and Standard FT. Training budgets are matched, but the direct arm encodes fewer documents and changes both the objective and negative comparisons; this is a pilot comparison of training recipes, not an isolated projection-only ablation.

Table 22: SciFact seed-42 test retrieval. All metrics are percentages. These are single-run values, not the multi-seed means in the main table.

Method	Selected step	nDCG@10	Hit@10	Recall@10
Pretrained	–	69.894	83.333	82.389
Standard FT	18	79.965	92.000	90.722
F	18	80.442	92.333	91.722
Direct U softplus	6	72.031	85.667	84.656

Let S_0 and F_0 denote queries successful and unsuccessful, respectively, under the pretrained full-space retriever. For a trained arm M,

$$\mathrm{F2S}_M = |F_0 \cap S_M|, \qquad \mathrm{S2F}_M = |S_0 \cap F_M|,$$

$$\mathrm{Hit@10}_M - \mathrm{Hit@10}_0 = \frac{\mathrm{F2S}_M - \mathrm{S2F}_M}{N}.$$

The last identity uses Hit as a fraction; multiply by 100 for percentage-point differences. There are $|F_0| = 50$ failures and $|S_0| = 250$ successes. All transitions below are recomputed from the saved per-query ranks using this same baseline and test-query order.

Table 23: Both directions of test-query transitions relative to pretrained retrieval. Recovery rate is F2S/50; damage rate is S2F/250.

Method	F2S	S2F	F2F	S2S	Recovery (%)	Damage (%)
Standard FT	29	3	21	247	58.0	1.2
F	27	0	23	250	54.0	0.0
Direct U softplus	9	2	41	248	18.0	0.8

The direct arm recovers nine failures and loses two prior successes, giving seven net additional successes ($+2.333$ Hit points). Its nDCG also improves over pretrained retrieval by 2.137 points. Therefore, the experiment does not support saying that direct U supervision cannot learn or necessarily causes collapse. Relative to F, however, it recovers 18 fewer baseline failures and incurs two additional losses, accounting for a 20-query deficit (6.667 Hit points); its nDCG is 8.411 points lower. The larger count difference is missed recovery, rather than widespread destruction of previous successes. Standard FT, in turn, recovers two more failures than F but loses three previous successes, illustrating why F2S alone is insufficient.

Taken together, the counterexample and pilot support a limited conclusion: projected-score optimization alone does not guarantee the full-space ordering used at inference, and this direct positive-only recipe is less effective than F in the observed run. The experiment does not isolate whether the difference arises from missing negative comparisons, loss saturation, residual-space effects, or their interaction. A multi-seed comparison with matched candidates and controlled objectives would be needed for a stronger causal or general empirical claim. F's full-space distributional target is consistent with the inference score, but it likewise provides no universal guarantee of improved held-out retrieval.

J Temperature Choice

We use $\tau = 0.05$ as a trade-off between retaining a non-degenerate supervision distribution and maintaining substantial probability mass on the positive document. Compared with $\tau = 0.02$, this setting increases the mean target entropy from 0.120 to 0.888 nats and the probability mass on the remaining candidates from 6.98% to 18.79%, while keeping the mean positive probability at 81.21%. At $\tau = 0.10$, the positive probability falls to 37.16%. Thus, 0.05 retains a richer distribution over alternatives while preserving strong positive emphasis. Entropy here measures distributional spread, rather than task-relevant information directly.

Table 24: Temperature-dependent statistics of the selective F target on archived Qwen–FiQA training batches. Probability and top-1 columns are percentages. Top-1 measures the selected positive's rank within the training candidates, not test retrieval. Entropy is averaged per occurrence, not computed from an averaged distribution.

τ	Positive prob.	Remaining mass	Entropy (nats)	Positive top-1
0.01	94.03	5.97	0.045	94.32
0.02	93.02	6.98	0.120	94.32
0.05	**81.21**	**18.79**	**0.888**	**94.31**
0.10	37.16	62.84	3.920	94.30