Dissociating Recoverable Features from Effective Interventions in Neural Reranking

Anonymous ACL submission

Abstract

LLM-based pointwise rerankers achieve strong retrieval performance, yet the internal computations behind their relevance decisions remain unclear. Prior analyses have shown that some LLM-based rerankers encode classical IR signals in their activations, but recoverability alone does not tell us which internal signals actually contribute to ranking behavior. We study this question in RankLLaMA-7B with a causal analysis pipeline that combines Partial Least Squares probing with neuron ablation. From early-layer MLP activations, we construct three candidate signals: a dominant lexical signal, a dominant score-predictive signal, and an aligned lexical-score signal. We then test whether the highest-loading neurons for each signal are causally important for ranking. The dominant lexical and score-predictive signals are weak intervention targets. The largest ranking drop comes instead from the aligned lexical-score signal: ablating its top neurons lowers NDCG@10 by about 0.053 on TREC DL19, localizes mainly to Layer 1, and replicates with similar magnitude on TREC DL20. These results identify a compact early-layer intervention-sensitive structure in a neural reranker and show that the internal signal with the largest effect on ranking is not the most obvious recoverable lexical or score signal.

1 Introduction

Fine-tuned LLM rerankers such as RankLLaMA substantially outperform traditional lexical baselines, turning an effectiveness gap into a mechanistic question. Prior results show BM25 Lucene at 0.506 NDCG@10 on TREC DL19, while fine-tuned LLM rerankers report up to 0.760 NDCG@10 on the same benchmark after task fine-tuning (Ma et al., 2024; Chowdhury et al., 2025; Nijasure et al., 2025). Appendix Table 6 summarizes these comparison points. This gap naturally raises the question of what information enables LLM rerankers to score passages so much more effectively than traditional lexical baselines. A natural first step is to ask whether fine-tuning causes the model to encode the lexical signals that have long mattered in IR, such as BM25, query-term overlap, and TF×IDF. Prior work by Chowdhury et al. (2025) probes a broad IR feature suite in RankLLaMA and finds that several lexical signals are recoverable, while others remain weak under their linear probing setup. That result establishes that classical IR structure is present in the model's hidden state. It does not yet tell us whether the same directions that support decoding are also the directions that the reranker relies on when it produces a score.

This distinction matters because recoverability and causal importance are not the same thing (Belinkov, 2022; Hewitt and Liang, 2019; Elhage et al., 2021; Meng et al., 2022). A feature may be readable from hidden activations without being part of the internal computation that most strongly shapes the model's ranking decision. The central question in this paper is therefore whether the internal signals that support feature recovery also lie on the path to ranking behavior. To answer that question, we combine readout and intervention. We first use Partial Least Squares (PLS) probing to identify low-dimensional lexical and score-related signals in MLP activations. PLS (Wold et al., 1985) is useful here because it preserves covariance with a chosen IR target, whereas PCA preserves only high-variance activation structure. We then test those signals causally by ablating the highest-loading neurons associated with them and measuring the resulting change in ranking quality.

The main comparison centers on three candidate signals extracted from the early half of RankLLaMA, that is, Layers 0 to 15. One summarizes the dominant lexical subspace and is denoted q_{lex}. A second summarizes the dominant score-predictive subspace and is denoted q_{score}. The third, q_{star}, is

the lexical signal most strongly aligned with the score subspace. These three candidates let us compare a simple lexical hypothesis, a simple score hypothesis, and an overlap hypothesis about which signal has the largest effect on ranking.

The main result is a clear dissociation. The dominant lexical and score directions are both easy to identify, but they are weak intervention targets. The substantial ranking damage comes instead from q_{star}, and nearly all of that effect is concentrated at Layer 1. In other words, the strongest intervention target in the base model is not the most obvious lexical direction or the most obvious score direction. It is a small early direction where lexical structure and score structure overlap.

The scope of the paper is intentionally narrow. The claim is about how one family of lexical signals is organized inside RankLLaMA-7B under the present intervention setup. We do not claim a full mechanistic decomposition of ranking, a universal cross-architecture mechanism, or a clean semantic relevance direction.

Cross-model reruns (Section 5.5) provide additional context. Under the same direct-construction pipeline, RankLLaMA3 and RankPythia shift the strongest effect toward q_{lex}, while RankMistral shows no comparably large intervention effect. We include these reruns to show that large intervention effects can remain concentrated in early layers while differing in their exact form across architectures. We will release the code used for the experiments and analyses upon acceptance.

2 Related Work

This paper sits at the intersection of neural reranking, probing-based representation analysis, and intervention-based interpretability. On the retrieval side, recent work has established RankLLaMA as a strong LLaMA-based pointwise reranker in a multi-stage retrieval pipeline (Ma et al., 2024). The closest starting point for the present paper is the probing study of Chowdhury et al. (2025), which probes a broad MSLR-style feature suite in RankLLaMA with ridge-regularized linear probes. Their results recover several query-coverage and TF×IDF-related signals, while BM25 and stream length remain weak in that setup. That result is important because it shows that some classical IR structure is recoverable from hidden activations, but it leaves open a different question: whether the directions that make those features easy to decode are also the directions that matter most for the model's ranking behavior.

That distinction follows a broader lesson from the probing literature. Probing is useful for showing what information is linearly recoverable from a model's representations (Belinkov, 2022; Conneau et al., 2018; Hewitt and Liang, 2019), but recoverability alone does not establish behavioral importance. A signal can be readable from hidden activations and still be only a correlated by-product of some broader latent computation. This motivates the move from descriptive readout to intervention. In that respect, our perspective is closer to causal and mechanistic work that asks whether changing an internal direction or computational substructure changes the model's output (Elhage et al., 2021; Meng et al., 2022; Geiger et al., 2022; Nanda et al., 2023; Conmy et al., 2023; Wang et al., 2023). It is also close to recent direction-based representation studies that identify low-dimensional activation subspaces with supervised methods and then test them causally. Heinzerling and Inui (2024), for example, use PLS to identify property-encoding directions in language models and intervene along those directions. Our use of PLS is closest to that line of work, although our setting is reranking and our intervention takes the form of neuron ablation. Within neural IR, Lu et al. (2025) show that a cross-encoder implements a relevance computation that resembles a semantic variant of BM25. That work is especially relevant here because it connects mechanistic analysis to classical IR structure, but it focuses on reconstructing a relevance pathway in a cross-encoder rather than testing feature-aligned neuron interventions in a reranking LLM.

Within neural IR, prior analyses have examined attribution in BERT rerankers (Jiang et al., 2021), evidence modeling for neural retrieval (Akkalyoncu Yilmaz et al., 2019), and sparse lexical retrieval architectures (Formal et al., 2021). Our paper is narrower than those lines of work in one important way. We do not try to explain reranking in general. Instead, we ask whether a small set of classical lexical IR signals, once decoded from RankLLaMA, identify the intervention targets with the largest ranking effect. This is the gap the paper addresses: whether feature-aligned directions are actually on the path to the final ranking score.

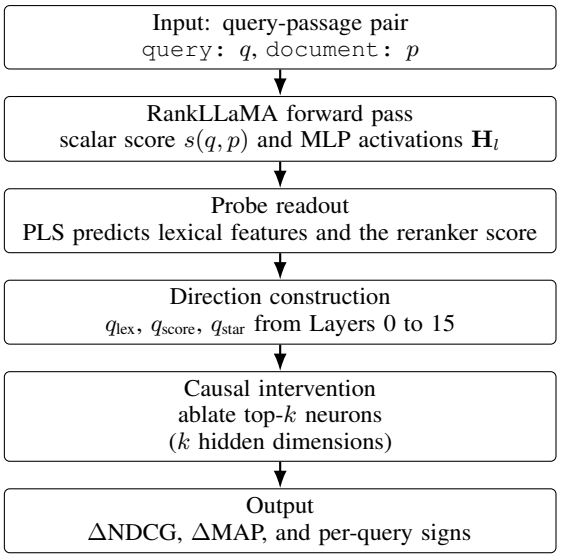

Figure 1: Method overview. The analysis first reads out lexical and score-aligned directions from MLP activations, then tests those directions by ablating the highest-loading k hidden dimensions during reranking.

3 Method

The method has two stages. First, we estimate directions that support lexical-feature recovery from hidden activations. Second, we ablate neurons associated with those directions during reranking and measure the change in ranking quality. Figure 1 summarizes the flow.

3.1 Model Activations

Each example is a query-passage pair (q, p). RankLLaMA is used as a pointwise reranker (Ma et al., 2024). For each query-passage pair, we pass `query:` q and `document:` p to the tokenizer as a text pair, which is then converted into a single model input sequence. The model produces a scalar reranking logit $s(q, p)$, and for a fixed query, candidate passages are ranked by this score. The same score is used as the target for score-aligned probes and, after reranking the same candidate set, for the downstream NDCG and MAP comparison.

We analyze the layerwise MLP outputs of RankLLaMA-7B. For each query-passage pair and transformer layer l, we summarize the tokenwise MLP output into one hidden vector for the pair, yielding an activation matrix $\mathbf{H}_l \in \mathbb{R}^{n \times d}$, where n is the number of query-passage pairs and $d = 4096$. The causal interventions are applied to the same MLP output space during the forward pass, before the final scalar score is produced. This gives a common representation space for comparing feature recoverability with intervention effect.

3.2 Feature Probes

The probing stage estimates which lexical IR signals are recoverable from these activations. For each feature-layer pair, we fit a partial least squares (PLS) regression probe from $\mathbf{H}_l$ to the scalar feature value and report held-out R^2. We use $m = 32$ components. PLS provides a supervised low-dimensional subspace whose components are selected to preserve covariance with the target feature. As an unsupervised comparison, we use the top-m PCA directions of $\mathbf{H}_l$, which have the same dimensionality but are selected only by activation variance. This comparison tests whether feature-aligned structure is concentrated in target-linked subspaces rather than in generic high-variance directions.

The probe results are used as evidence of feature presence in the representation. They are not the basis of the causal claim. The probe split is over query-passage pairs, so we treat probe R^2 as a recoverability measure within the evaluated topic distribution, not as a separate query-generalization result.

Figure 2 makes the next step explicit. The black path is descriptive: it traces what can be read out from hidden activations. The red path is causal: a selected direction determines a neuron set, that set is ablated in the hidden state, and the resulting change is measured at the score and ranking levels.

3.3 Intervention Direction Construction in Band A

Candidate intervention directions are extracted from the early half of the model, which we denote by Band A and define as Layers 0 to 15. This choice provides a broad early region in which lexical and score-related structure can be compared before later layers are more tightly coupled to the final scoring computation. Appendix E gives the motivation for this band.

The construction compares three hypotheses about what the largest early intervention effect might be. If the intervention-sensitive structure is primarily lexical, then q_{lex} should be the most damaging direction. If it is primarily tied to the model's internal scoring axis, then q_{score} should dominate. If the decisive signal lies in the overlap between lexical structure and score-relevant structure, then q_{star} should be the stronger intervention target.

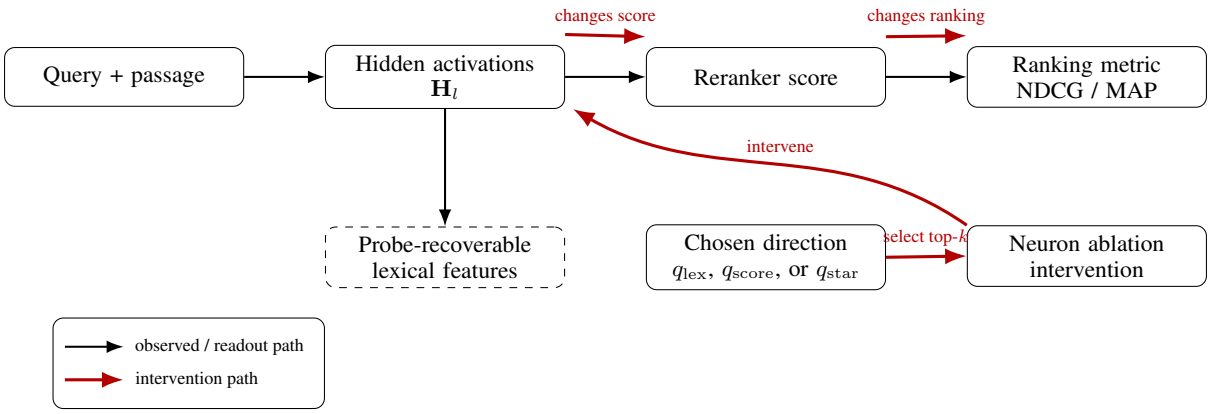

Figure 2: Causal schematic of the method. Probing asks what can be read out from hidden activations. Ablation asks whether a selected direction, realized by its top-k neurons, changes the score and downstream ranking metrics.

In short, q_{lex} is the main pooled lexical direction, q_{score} is the main pooled score direction, and q_{star} is the lexical direction most aligned with the score subspace.

For each layer in Band A, we fit feature-specific PLS bases for the lexical targets and score-predictive PLS bases for the reranker score. The per-layer lexical bases are then pooled into a band-level lexical basis $\mathbf{Q}^A_{\text{lex}} \in \mathbb{R}^{d\times r}$, and the per-layer score bases are pooled into a band-level score basis $\mathbf{Q}^A_{\text{score}} \in \mathbb{R}^{d\times r}$. From each pooled basis, we take the dominant left singular vector, that is, the unit hidden-space direction most aligned with the combined per-layer PLS bases on average. These give q_{lex} and q_{score}. Thus q_{lex} is aligned to the dominant pooled lexical subspace induced by the PLS fits across the lexical targets, rather than to any single feature. Concretely, q_{lex} is the first left singular vector of the pooled lexical PLS basis over Band A, not the coefficient vector of one probe. The third direction, q_{star}, is the lexical-space unit vector with maximal alignment to the score basis, obtained from the top singular vector of $(\mathbf{Q}^A_{\text{lex}})^\top \mathbf{Q}^A_{\text{score}}$. For each direction, neurons are ranked by squared loading, and the top-k coordinates are ablated. The intervention therefore uses the direction only to rank which hidden dimensions to mask; it does not apply a weighted projection or steering update along that direction. Here k denotes the number of hidden dimensions removed.

3.4 Neuron Ablation

Given a set of k ablated neuron indices $\mathcal{S}$ identified from one of these directions, we register the same forward hook on every MLP layer in the selected band so that the same hidden dimensions are zeroed throughout the early stack:

$$\mathbf{h}^{(l)} \leftarrow \mathbf{h}^{(l)} \odot \mathbf{m}_{\mathcal{S}}, \qquad l \in \text{Band A},$$
$$(\mathbf{m}_{\mathcal{S}})_i = \begin{cases} 0 & i \in \mathcal{S}, \\ 1 & \text{otherwise.} \end{cases}$$

Here $\odot$ denotes elementwise (Hadamard) multiplication, and $\mathbf{m}_{\mathcal{S}}$ is a binary mask that zeros out the k ablated neurons and leaves the rest unchanged. All subsequent computations use the activations produced after ablation. In the main causal test, we apply each direction across the full early band and evaluate the resulting change in ranking quality. We test $k \in \{16, 32, 64, 128\}$ for each intervention direction. A direction is treated as causally important when ablating its highest-loading neurons produces a reliable drop in ranking performance.

Random null distribution. For each ablation we generate $N = 20$ random neuron sets of the same size k and compute their mean and standard deviation of ΔNDCG. We report the corresponding z-score,

$$z = \frac{\Delta\text{NDCG}_{\text{dir}} - \mu_{\Delta\text{NDCG}_{\text{random}}}}{\sigma_{\Delta\text{NDCG}_{\text{random}}}},$$

but we treat z as secondary to the observed effect size and its confidence interval because the random null is small.

3.5 Evaluation

All main causal experiments are evaluated by NDCG@10 against TREC Deep Learning 2019 (DL19) relevance judgments (Craswell et al., 2020) (43 queries, grades 0 to 3). We report $\Delta\text{NDCG} = \text{NDCG}_{\text{intv}} - \text{NDCG}_{\text{base}}$ and paired per-query bootstrap confidence intervals (2000 resamples). DL20

is used as a held-out benchmark-year replication check for the base-model q_{star} intervention.

4 Experimental Setup

DL19 is used for the main base-model causal test. DL20 is used as a benchmark-year replication check for the same intervention directions.

Model. RankLLaMA-7B (v1; `castorini/rankllama-v1-7b-lora-passage`), loaded via HuggingFace Transformers (Wolf et al., 2020) with the base tokenizer from `meta-llama/Llama-2-7b-hf`. The model has 32 transformer layers, hidden dimension $d = 4096$, and LoRA rank $r = 8$ on attention projections. The MLP weights remain those of the pretrained Llama-2-7B checkpoint.

DL19 evaluation. TREC Deep Learning 2019 passage ranking (Craswell et al., 2020): 43 judged queries with top-50 RepLLaMA candidate passages, giving $n = 2{,}150$ query-passage pairs. The baseline (no intervention) is NDCG@10 $= 0.7629$, MAP $= 0.8534$.[1]

DL20 replication. TREC Deep Learning 2020 (DL20) passage ranking (Craswell et al., 2021): 54 judged queries with top-200 RepLLaMA candidate passages, giving $n = 10{,}800$ pairs. Using the same RankLLaMA model, the baseline is NDCG@10 $= 0.7502$, MAP $= 0.7406$. The DL20 experiment reuses the Band A directions extracted on DL19 and tests whether the same intervention directions produce similar effects in a separate TREC year.

Query split check. DL19 and DL20 topics correspond to MS MARCO eval queries. In the query files used here, neither DL19 nor DL20 overlaps MS MARCO train or dev queries by query id or exact normalized query text, and the DL19 and DL20 topic sets are mutually disjoint.

IR feature computation. Following Chowdhury et al. (2025), who probe a broad MSLR-style feature suite, we focus on four lexical signals: BM25, covered query term ratio (CQTR), mean TF×IDF, and stream length. In their setup, CQTR and TF×IDF-related signals were among the clearer positive cases, whereas BM25 and stream length were weak. Table 1 shows that all four are recoverable under our denser PLS probe. These four are drawn from a broader 19-feature lexical suite and give a compact comparison set that spans both earlier-positive and earlier-weak probe cases. We include this mix deliberately so that the later intervention analysis can compare recoverability against causal importance across both earlier-positive and earlier-weak features. BM25 uses per-passage statistics precomputed over MS MARCO with IDF values from the full passage collection. The other three are computed directly from raw passage text. Features are standardized before PLS fitting but retained in original scale for descriptive correlation analyses.

[1] NDCG is computed with full-pool IDCG normalization, i.e., the ideal DCG is computed over all judged documents in the TREC assessment pool rather than only the top-50 candidate set.

Table 1: PLS vs. PCA probe R^2 at $m = 32$. Best layer per feature shown. The mean is computed over all feature-layer pairs in the broader lexical suite, not only over the four representative rows shown here.

Feature	Layer	PLS R^2	PCA R^2
CQTR	29	0.982	0.588
Mean TF×IDF	29	0.903	0.606
Stream length	0	0.961	0.821
BM25	29	0.847	0.574
Mean across all feature-layer pairs		0.773	0.576

PLS configuration. We use $m = 32$ directions for per-layer probing. Band-level direction construction uses rank $r = 4$ pooled lexical and score bases, which are then reduced to the rank-1 directions q_{lex}, q_{score}, and q_{star} for top-k neuron selection.

Supplementary analyses. Vocabulary projection, large-scale projection over 200,000 MS-MARCO pairs, and a targeted diagnostic benchmark are reported in the appendix.

5 Results

5.1 IR Features Are Recoverable

We first test feature recoverability. Table 1 shows representative PLS probing results at $m = 32$. Across all feature-layer pairs, mean probe $R^2 = 0.773$ for PLS versus 0.576 for matched PCA. Feature-aligned subspaces therefore preserve substantially more IR-relevant information than untargeted high-variance directions of the same dimension.

These results establish the representational premise of the paper, but they do not establish

Table 2: Neuron ablation results on TREC DL19 (baseline NDCG@10 = 0.763). q_{lex}: dominant lexical direction. q_{score}: dominant score direction. q_{star}: lexical direction maximally aligned with the score basis. k is the number of hidden dimensions ablated.

Direction	k	ΔNDCG	95% CI	ΔMAP	z
q_{lex}	64	−0.001	[−0.008, +0.005]	+0.002	−0.58
q_{lex}	128	−0.002	[−0.014, +0.009]	−0.001	+0.07
q_{score}	64	−0.002	[−0.010, +0.006]	−0.005	−0.23
q_{score}	128	−0.010	[−0.027, +0.007]	−0.014	−0.36
q_{star}	**64**	**−0.052**	**[−0.078, −0.027]**	**−0.060**	**−18.6**
q_{star}	**128**	**−0.053**	**[−0.079, −0.027]**	**−0.055**	**−12.1**

Table 3: Projection of intervention directions onto the Layer 1 score-basis directions. q_{star} lies mostly in score PC2+PC3 (85% combined), while q_{score} is almost entirely score PC4 (98%).

Direction	Score PC1	Score PC2	Score PC3	Score PC4	$\|\text{proj}\|^2$
q_{score}	0%	2%	0%	98%	1.00
q_{star}	8%	51%	34%	7%	1.00
q_{lex}	0%	30%	68%	2%	0.36

All percentages show share of each direction's projection within the score subspace. $\|\text{proj}\|^2$: fraction of the unit vector that falls within the score subspace.

causal importance. They show that the four lexical signals are recoverable from hidden state. The DL19 topics are disjoint from MS MARCO train/dev queries, but the probe split itself is pair-disjoint, not query-disjoint, so the probe scores should not be interpreted as a strong query-level generalization result.

Peak decoding layer is not the main causal layer. The peak probe layers are late for BM25, CQTR, and mean TF×IDF, and early for stream length. That pattern becomes more informative once we turn to the causal results, because the largest intervention effect is concentrated much earlier, at Layer 1. Taken together, these results suggest that, within the present intervention setup, the layer where a feature is easiest to decode need not coincide with the main intervention-sensitive layer.

5.2 A Lexical Direction Aligned with the Score Subspace Has the Largest Intervention Effect

We next test whether the directions identified in the probing stage are also effective intervention directions. Table 2 reports the DL19 ablation results for top-k neurons selected by the three Band A directions.

The dominant lexical direction is weak: ablating q_{lex} changes NDCG by at most about -0.002. The dominant score direction is also weak: even at $k = 128$, q_{score} reaches only -0.010, with a confidence interval that still includes zero. q_{star} behaves differently. Ablating q_{star} drops NDCG@10 by about -0.053 at both $k = 64$ and $k = 128$, with clear separation from zero. It is the only tested signal whose confidence interval excludes zero at both values of k. The z-scores are consistent with this pattern, but the main evidence is the effect size and its bootstrap interval. Among the three candidate directions, q_{star} is the only one that produces a large and stable ranking drop.

Per-query consistency. The q_{star} effect is broad rather than outlier-driven (Appendix Table 8). At $k = 64$, 31 of 43 queries worsen, 8 improve, and 4 are unchanged. At $k = 128$, 30 worsen, 8 improve, and 5 are unchanged. By contrast, q_{lex} and q_{score} remain centered near zero.

Interpretation of q_{star}. q_{star} is not simply a union of lexical and score subspaces. It is the lexical-space signal that maximally overlaps the score basis. The ablation result therefore gives a more specific picture of the reranker's early computation: the dimensions with the largest intervention effect are not the dominant lexical dimensions or the dominant score dimensions taken separately, but dimensions shared by both.

Score-basis component analysis. To understand where each direction sits within the rank-4 score subspace, we decompose the three directions into the score-basis PCs (Table 3). q_{star} lives in score PC2 and PC3 (85% combined). q_{score} is almost entirely score PC4 (98%) and is causally weak. A rank sweep over individual score-basis components confirms that score PC2 carries the causal signal: ablating score PC1 neurons is inert (ΔNDCG $= +0.003$ at k=64), while score PC2 is strongly destructive (ΔNDCG $= -0.056$, z=−15.7). The full sweep is in Table 4.

5.3 The Base-Model Causal Effect Is Concentrated at Layer 1

We then localize the q_{star} effect. Figure 3 (left) reuses the same Band A q_{star} neuron set, but applies the ablation to one early layer at a time. Almost all of the effect is concentrated at Layer 1.

At $k = 64$, ablating q_{star} only at Layer 1 gives ΔNDCG $= -0.0491$, which is 93.8% of the full-band effect (-0.0524). At $k = 128$, Layer 1 alone

Table 4: Score-basis component sweep results. Baseline NDCG@10 = 0.763. Band A ablation (Layers 0 to 15). $n_{\text{controls}} = 10$.

Spec	k	ΔNDCG	ΔMAP	ctrl mean	ctrl std	z
score PC1	64	+0.003	−0.002	+0.000	0.003	+0.94
score PC1	128	+0.000	−0.003	−0.004	0.003	+1.38
score PC2	64	−0.056	−0.059	+0.001	0.004	−15.7
score PC2	128	−0.063	−0.066	−0.003	0.006	−10.3
score PC3	64	−0.001	−0.004	−0.002	0.005	+0.23
score PC3	128	−0.057	−0.060	−0.003	0.003	−17.7
score PC4	64	−0.001	−0.003	−0.001	0.003	+0.02
score PC4	128	−0.053	−0.059	−0.003	0.005	−10.8

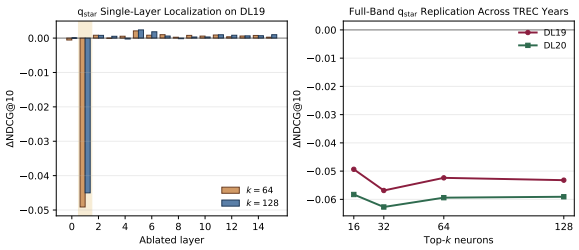

Figure 3: Localization and replication of the q_{star} effect, reported as ΔNDCG@10. Left: single-layer reapplication of the Band A q_{star} neuron set on DL19 shows that nearly all of the damage is concentrated at Layer 1 for both $k = 64$ and $k = 128$. Right: the full-band q_{star} ablation produces similar damage on DL19 and DL20.

gives $\Delta\text{NDCG} = -0.0450$, or 84.6% of the full-band effect (-0.0532). Every other single early layer is near zero.

Same-layer and exclusion controls. Two controls test the specificity of this localization. First, applying q_{lex} or q_{score} only at Layer 1 produces $\Delta\text{NDCG} \approx 0$ at both k values. Second, ablating Band A without Layer 1, that is, Layer 0 together with Layers 2 to 15, also leaves NDCG near zero. The result is therefore not that untargeted early-layer damage is harmful. It is that the main intervention effect is largely concentrated in a specific Layer 1 direction.

Layer 1 geometry. An independent PCA audit of the saved DL19 activation bank confirms why Layer 1 is the localization site. Within Band A, Layer 1 is uniquely collapsed: the top 4 PCs capture 99.99% of its variance, versus 47.4% at Layer 0 and only 20.2%–28.7% across Layers 2 to 15 (Figure 4). Projecting out just the top Layer 1 PC drops NDCG@10 by -0.349; by 16 PCs the drop is -0.369.

The matched-null comparison is an important caveat: removing the top 16 Layer 1 PCs drops

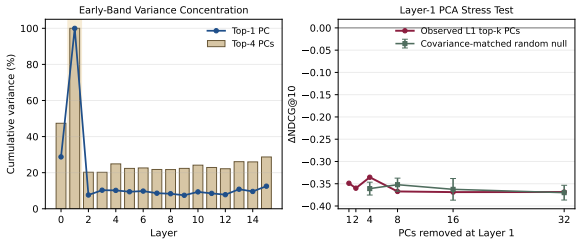

Figure 4: Independent PCA stress test over the saved DL19 activation bank. Left: cumulative explained variance in Band A shows that Layer 1 is the only early layer whose top 4 PCs capture nearly all variance; dots mark the top-1 PC mass. Right: ΔNDCG@10 after projecting out top Layer 1 PCs versus matched covariance-weighted random Layer 1 subspaces; both are highly destructive once dimensionality is matched.

NDCG by -0.369, but a covariance-weighted random 16-D subspace already averages -0.363. Layer 1 is both structurally distinctive and broadly fragile. The specific evidence for q_{star} comes from its selectivity: it strongly outperforms same-size random ablations ($z = -18.6$ to -30.7), whereas q_{lex} and q_{score} remain weak at Layer 1.

5.4 The Base-Model Effect Replicates on DL20

We evaluate the same Band A directions on DL20 to test whether the effect persists in a separate TREC year. The right panel of Figure 3 shows a similar q_{star} effect. For q_{star}, the effect is $\Delta\text{NDCG} = -0.059$ (95% CI $[-0.099, -0.019]$) at $k = 64$ and $\Delta\text{NDCG} = -0.059$ (95% CI $[-0.100, -0.018]$) at $k = 128$, with corresponding ΔMAP values of -0.117 and -0.115.

The dissociation also persists on DL20: q_{lex} remains weak and q_{score} remains weak, with confidence intervals that include zero. Per-query consistency again supports a broad effect rather than a few outliers. Taken together, the DL19 and DL20 experiments support a stable base-model claim about which tested signal has the largest effect in the base model.

Appendix analyses of vocabulary projection, large-scale projection, and a targeted diagnostic benchmark provide additional interpretive context. Section 5.5 below extends these results to three further reranker architectures.

5.5 The Dominant Intervention Target Is Architecture-Specific

Table 5 summarizes direct cross-model results under the same construction pipeline. RankL-

Table 5: Cross-model ablation results under the same direct-construction pipeline. The dominant intervention target varies by architecture. For RankMistral, "largest-magnitude spec" denotes the 4-layer window with largest effect magnitude for q_{star}.

Model	Direction / Spec	k	ΔNDCG	z
RankLLaMA2	q_{lex}, Band A	128	−0.002	+0.07
	q_{star}, Band A	64	−0.052	−18.6
	q_{star}, L1 only	64	−0.049	−30.7
RankLLaMA3	q_{lex}, Band A	128	−0.313	−45.0
	q_{lex}, L0 only	128	−0.319	−81.8
	q_{star}, Band A	128	−0.025	−1.22
RankPythia	q_{lex}, Band A	128	−0.235	−75.7
	q_{star}, Band A	64	−0.015	−2.19
RankMistral	q_{star}, largest-magnitude spec	64	−0.004	−1.36

LaMA3, RankPythia, and RankMistral are fine-tuned rerankers introduced in Nijasure et al. (2025) following the same LoRA pointwise training setup as RankLLaMA-7B (Ma et al., 2024). The dominant intervention target is not stable across architectures. In RankLLaMA-7B, the largest effect comes from q_{star}. In RankLLaMA3 and RankPythia, the largest effect shifts to q_{lex}. RankMistral shows no comparably large intervention effect under this construction.

The cross-model reruns show that the dominant intervention target depends on the model.

6 Discussion

The paper makes three linked contributions. First, it introduces a reranker-focused causal analysis pipeline that connects feature recovery to intervention-based testing. Second, it shows a concrete empirical dissociation: the dominant lexical signal and the dominant score-predictive signal are both recoverable, yet neither is the strongest intervention target. Third, it identifies a more specific structural result in the base model: the largest intervention effect comes from a compact early signal aligned with both the lexical and score subspaces and concentrated at Layer 1.

This result sharpens what probing can tell us in neural reranking. Probing is useful for establishing that classical IR structure is present in the model's hidden state, but it does not by itself identify the internal signal with the largest effect on ranking behavior. In RankLLaMA-7B, the most intervention-sensitive early signal is not the dominant lexical axis or the dominant score axis taken separately.

The interpretation remains limited. The evidence supports a compact intervention-sensitive structure in RankLLaMA-7B, not a full mechanistic decomposition of relevance scoring and not a universal reranker mechanism. Appendix analyses give additional context: semantic PLS at Layer 1 is weak, vocabulary projection looks more template- or format-associated than semantic, and the targeted benchmark plus manual audit point toward a coarse positive-evidence or answer-likeness signal rather than a clean semantic relevance detector.

The DL20 replication shows that the base-model result is not confined to a single benchmark year. The cross-model reruns extend the picture in a different way: when large intervention effects appear, they tend to remain early-layer, but their exact dominant signal can vary across reranker families. That makes the main finding more specific rather than less useful. The rank-4 bottleneck at Layer 1 and the geometry of q_{star} provide concrete structural targets for future work. The present results make it possible to ask whether this organization is preserved under domain fine-tuning, whether it can support compression or pruning, and whether analogous early bottlenecks appear in rerankers trained with different objectives.

7 Conclusion

We test whether probe-recoverable lexical structure identifies the internal signals that matter most for ranking in RankLLaMA-7B. It does so only partially. The dominant lexical and score-predictive signals are recoverable, but the largest intervention effect comes from q_{star}, an early signal aligned with both lexical and score-related structure. That effect is concentrated at Layer 1 and replicates on DL20. Taken together, the results support a concrete picture of the base model's early ranking computation: the internal signal with the largest effect on ranking is compact, early, and not identical to the most obvious recoverable lexical or score signal. Understanding exactly what computation this aligned signal implements remains open.

8 Limitations

This paper makes a narrow causal claim, and several limitations matter for how far that claim should be taken.

- **The probing result is descriptive, not definitive.** The DL19 topics are disjoint from the MS MARCO train/dev query files used here, which removes a train/dev query-overlap concern. The recoverability analysis still uses

pair-disjoint splits on a small judged query set, so we read the probe R^2 results as evidence that these signals are present in the activations, not as a strong query-level generalization claim.

- **Direction extraction and DL19 evaluation use the same benchmark pool.** The DL20 replication is therefore the main external check on the discovery-year result.
- **Random-null z-scores are secondary evidence.** The main paper relies primarily on effect size and bootstrap confidence intervals. The same-size random-neuron null uses only 20 draws, so the reported z-scores should be read as coarse specificity summaries, not precise tail estimates.
- **Interpretation is weaker than the causal evidence.** The appendix-level semantic PLS, vocabulary projection, and targeted benchmark analyses suggest a coarse positive-evidence or answer-likeness signal, but they do not identify a clean semantic relevance feature or a full mechanistic decomposition.
- **Cross-model generalization remains exploratory.** The strongest evidence is for RankLLaMA-7B on DL19 and DL20. The cross-model reruns in Section 5.5 are informative but mixed, so we do not claim a universal reranker mechanism.

Acknowledgments

We used large language model (LLM) assistance during preparation of this manuscript. Specifically, LLM tools were used for language polishing and wording refinement, occasional coding and debugging support, and for reviewing and brainstorming presentations of ideas. In some cases, LLM assistance was used to help draft and refine portions of the appendix and method exposition based on implemented code and directed author prompts. LLM tools were not used to originate the research question or study design, and design experiments.

References

Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3490–3496, Hong Kong, China. Association for Computational Linguistics.

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

Tanya Chowdhury, Atharva Nijasure, and James Allan. 2025. Probing ranking llms: A mechanistic analysis for information retrieval. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval*, ICTIR '25, pages 336–346, Padua, Italy. Association for Computing Machinery.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. In *Proceedings of TREC*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR.

Benjamin Heinzerling and Kentaro Inui. 2024. Monotonic representation of numeric attributes in language

models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195, Bangkok, Thailand. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Zhiying Jiang, Raphael Tang, Ji Xin, and Jimmy Lin. 2021. How does BERT rerank passages? an attribution analysis with information bottlenecks. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 496–509, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Meng Lu, Catherine Chen, and Carsten Eickhoff. 2025. Pathway to relevance: How cross-encoders implement a semantic variant of BM25. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25525–25547, Suzhou, China. Association for Computational Linguistics.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. page 2421–2425.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. volume 35, pages 17359–17372.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability.

Atharva Nijasure, Tanya Chowdhury, and James Allan. 2025. How relevance emerges: Interpreting lora fine-tuning in reranking llms. *arXiv preprint arXiv:2504.08780*.

Tao Qin and Tie-Yan Liu. 2013. Introducing letor 4.0 datasets. *arXiv preprint arXiv:1306.2597*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Herman Wold, S Kotz, and NL Johnson. 1985. *Encyclopedia of statistical sciences*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Reranker Effectiveness Context

Table 6 summarizes the prior effectiveness context used to motivate the causal question. Panel A reports the BM25 and fine-tuned LLM reranker comparison from Chowdhury et al. (2025). Panel B reports prior DL19 LoRA component ablations including Mistral-7B (Nijasure et al., 2025). The panels come from different prior experiments, so they should be read as context for the effectiveness gap rather than as a new controlled benchmark.

B Feature Scope

The main paper focuses on four representative lexical features: BM25, covered query term ratio, mean TF×IDF, and stream length. The experimental code also implements a broader 19-feature MSLR-style lexical suite (Qin and Liu, 2013), including covered-term count, term-frequency aggregates, stream-length-normalized term-frequency aggregates, and TF×IDF aggregates. We do not treat the current paper as evidence about all 19 features. The broad suite is useful for audits and reruns, but the main causal claim is already driven by the base-model q_{star} result.

For the four representative features used in the main paper, the computation follows the same lexical feature setup as the prior RankLLaMA probing study of Chowdhury et al. (2025). Covered query term ratio (CQTR) is the fraction of query terms that appear in the passage, that is, the number of distinct matched query terms divided by the number of query terms. Mean TF×IDF is computed by taking, for each query term, its term frequency in the passage multiplied by its collection IDF, then averaging those values across the query terms. Stream length is the passage length used by the lexical feature pipeline, i.e., the number of passage terms after the same preprocessing used for the feature computation. BM25 is standard passage-level BM25 using collection-level IDF values from the MS MARCO passage collection together with the corresponding per-passage term statistics. In our pipeline, BM25 uses precomputed per-passage statistics, while the other three features are computed directly from raw passage text and then standardized only for probe fitting.

Table 6: Prior effectiveness context for the paper's motivating question. Panel A reports BM25 and fine-tuned LLM reranker results from Chowdhury et al. (2025); their percentage-form scores are converted here to the 0–1 scale used elsewhere in this paper. Panel B reports DL19 NDCG@10 from LoRA component ablations in Nijasure et al. (2025).

A. BM25 and fine-tuned LLM rerankers				
Model	**Base model**	**DEV MRR@10**	**DL19 NDCG@10**	**DL20 NDCG@10**
BM25	Lucene	0.167	0.506	0.480
RankLLaMA2-7B	LLaMA2	0.449	0.756	0.774
RankLLaMA2-13B	LLaMA2	0.452	0.760	0.779
RankLLaMA3-8B	LLaMA3.1	0.428	0.743	0.724
RankPythia-6.9B	Pythia	0.429	0.760	0.715
B. DL19 LoRA component effects				
Base model	**No LoRA updates**	**MLP only**	**MHA only**	**MHA + MLP**
LLaMA3.1-8B	0.1855	0.4341	0.5987	0.7655
Mistral-7B	0.1450	0.4390	0.5164	0.6891
Pythia-6.9B	0.1895	0.6262	0.7009	0.7569

C IR Evaluation Details

We evaluate ranking quality with NDCG@10 and MAP under the standard TREC Deep Learning passage-ranking setup. NDCG uses full-pool IDCG normalization, meaning that the ideal DCG is computed over all judged documents in the TREC assessment pool rather than only the reranker's candidate set. This matches standard TREC evaluation practice and avoids defining the ideal ranking only within the top-ranked candidates supplied to the reranker.

For a query with ranked documents $d_1, \ldots, d_k$ and graded relevance labels $g_1, \ldots, g_k$, discounted cumulative gain at cutoff k is

$$\mathrm{DCG@}k = \sum_{i=1}^{k} \frac{g_i}{\log_2(i+1)}.$$

Normalized discounted cumulative gain divides this quantity by the ideal DCG for the same query:

$$\mathrm{NDCG@}k = \frac{\mathrm{DCG@}k}{\mathrm{IDCG@}k}.$$

In this paper, IDCG@10 is computed from the full set of judged documents in the TREC pool for that query, sorted by decreasing relevance grade, rather than only from the reranker's candidate set.

Mean average precision (MAP) is computed from binary relevance judgments derived from the TREC labels, where any document with positive relevance grade is treated as relevant. For a query with ranked binary relevance indicators $r_1, \ldots, r_n \in \{0, 1\}$, average precision is

$$\mathrm{AP} = \frac{1}{\sum_{i=1}^{n} r_i} \sum_{i=1}^{n} r_i \cdot \mathrm{Prec@}i,$$

where $\mathrm{Prec@}i$ is the precision among the top i ranked documents. MAP is the mean of AP over the evaluated queries.

D Compute and Release Details

The experiments used one NVIDIA A100 GPU primarily for inference-time reranking and causal-effect measurement. PLS fitting and the associated analysis code were run on CPUs.

E Why Use Band A?

RankLLaMA-7B has 32 transformer layers. We define Band A as Layers 0 to 15, the early half of the model. This is a deliberately broad search region rather than a claim that all useful ranking computation happens before Layer 16. The purpose is to capture early lexical and score-forming structure while avoiding a layer-by-layer search over the full network as the primary discovery procedure.

There are two reasons this choice is appropriate for the main causal test. First, the hypothesis being tested is about compact early directions that connect lexical structure to the reranker score. Second, the follow-up localization experiment shows that the Band A q_{star} effect is not diffuse across the whole band: when the same neuron set is applied one layer at a time, the effect is concentrated at Layer 1. The exclusion control, which ablates Band A without Layer 1, remains near zero. Thus Band A functions as a broad early discovery region, while the evidence itself identifies Layer 1 as the main site of the effect.

F PLS and Direction Construction Details

This section makes explicit the mathematics behind the probing and direction-construction code.

Per-layer notation. For a fixed layer l, let $\mathbf{H}_l \in \mathbb{R}^{n\times d}$ denote the matrix of MLP activations over n query-passage pairs, with one row per pair and hidden dimension $d = 4096$. For a scalar target feature f, let $\mathbf{y}_f \in \mathbb{R}^n$ be its value over the same pairs. Before probe fitting, each scalar target feature is standardized to zero mean and unit variance for numerical stability; the underlying IR features are otherwise computed in their usual form.

PLS probing. We use partial least squares regression (Wold et al., 1985) with m latent components. Let $\mathbf{X}^{(0)}$ be the centered activation matrix and let $\mathbf{y}^{(0)}$ be the standardized target. PLS is supervised: unlike PCA, it chooses directions using the target rather than activation variance alone. At component j, it selects a direction whose projected activation scores have high covariance with the current residual target,

$$\mathbf{w}_j = \arg\max_{\|\mathbf{w}\|_2=1} \mathrm{Cov}^2(\mathbf{X}^{(j-1)}\mathbf{w}, \mathbf{y}^{(j-1)}),$$
$$\mathbf{t}_j = \mathbf{X}^{(j-1)}\mathbf{w}_j.$$

The score $\mathbf{t}_j$ is then used to remove the explained component from both the activations and the target:

$$\mathbf{p}_j = \frac{(\mathbf{X}^{(j-1)})^\top \mathbf{t}_j}{\mathbf{t}_j^\top \mathbf{t}_j}, \qquad c_j = \frac{(\mathbf{y}^{(j-1)})^\top \mathbf{t}_j}{\mathbf{t}_j^\top \mathbf{t}_j},$$

$$\mathbf{X}^{(j)} = \mathbf{X}^{(j-1)} - \mathbf{t}_j\mathbf{p}_j^\top, \qquad \mathbf{y}^{(j)} = \mathbf{y}^{(j-1)} - c_j\mathbf{t}_j.$$

After m components, PLS gives a low-dimensional score matrix

$$\mathbf{T}_l = \mathbf{H}_l\mathbf{W}_l,$$

and the fitted probe can be written in the original hidden space as

$$\hat{\mathbf{y}}_f = \mathbf{H}_l\mathbf{B}_{f,l} + b,$$

where $\mathbf{B}_{f,l} \in \mathbb{R}^d$ is the PLS coefficient vector and b is the intercept. In the code, this is the object returned by `PLSRegression`, and the reported score is held-out probe R^2. The main probe table uses $m = 32$. Direction extraction uses the fitted PLS directions to form smaller rank-r lexical and score bases before constructing $\mathrm{q}_{\mathrm{lex}}$, $\mathrm{q}_{\mathrm{score}}$, and $\mathrm{q}_{\mathrm{star}}$.

Matched PCA baseline. For comparison, PCA computes the top-m variance directions of $\mathbf{H}_l$ without reference to the target. If $\mathbf{U}_{l,m} \in \mathbb{R}^{d\times m}$ contains those directions, then the matched PCA probe fits the same downstream regressor on $\mathbf{H}_l\mathbf{U}_{l,m}$. This comparison isolates whether feature-aligned directions retain more predictive information than untargeted high-variance directions of the same dimension.

Per-layer lexical and score bases. For each layer in the early band and for each lexical feature $f \in \mathcal{F}$, we fit a PLS model and retain a small number r of feature-aligned directions. We denote the resulting per-layer basis by

$$\mathbf{Q}_f^l \in \mathbb{R}^{d\times r}.$$

We do the same for the reranker score target $\mathbf{s} \in \mathbb{R}^n$, producing

$$\mathbf{Q}_{\mathrm{score}}^l \in \mathbb{R}^{d\times r}.$$

Operationally, these are the leading directions retained from the fitted PLS models for each target.

Band-level pooling. Let $A = \{0, \ldots, 15\}$ denote the early-layer band used in the paper. The band-level lexical basis is formed by concatenating the feature-specific bases across all lexical features and all layers in the band, then reducing the result to rank r:

$$\mathbf{Q}_{\mathrm{lex}}^A = \mathrm{Reduce}_r\left[\mathbf{Q}_f^l\right]_{l\in A,\, f\in\mathcal{F}} \in \mathbb{R}^{d\times r}.$$

The band-level score basis is defined analogously:

$$\mathbf{Q}_{\mathrm{score}}^A = \mathrm{Reduce}_r\left[\mathbf{Q}_{\mathrm{score}}^l\right]_{l\in A} \in \mathbb{R}^{d\times r}.$$

In the implementation, the reduction is an SVD-based low-rank summary of the concatenated per-layer directions.

Dominant and aligned directions. The dominant lexical direction $\mathrm{q}_{\mathrm{lex}}$ is the leading left singular vector of $\mathbf{Q}_{\mathrm{lex}}^A$, and the dominant score direction $\mathrm{q}_{\mathrm{score}}$ is the leading left singular vector of $\mathbf{Q}_{\mathrm{score}}^A$:

$$\mathbf{q}_{\mathrm{lex}} = \mathrm{TopLeftSV}(\mathbf{Q}_{\mathrm{lex}}^A),$$
$$\mathbf{q}_{\mathrm{score}} = \mathrm{TopLeftSV}(\mathbf{Q}_{\mathrm{score}}^A).$$

Here the leading left singular vector is the unit hidden-space direction

$$\arg\max_{\|\mathbf{q}\|_2=1} \|(\mathbf{Q}^A)^\top\mathbf{q}\|_2,$$

so it is the single direction that best summarizes the pooled basis in the original hidden dimension.

The aligned direction q_{star} is defined differently. It is the lexical-space unit vector whose image under the lexical basis is maximally aligned with the score basis. Writing

$$\mathbf{M} = (\mathbf{Q}^A_{\text{lex}})^\top \mathbf{Q}^A_{\text{score}} \in \mathbb{R}^{r\times r},$$

we take the top singular vector $\mathbf{u}_1$ of $\mathbf{M}$ and map it back to hidden space:

$$\mathbf{q}_{\text{star}} = \frac{\mathbf{Q}^A_{\text{lex}}\mathbf{u}_1}{\|\mathbf{Q}^A_{\text{lex}}\mathbf{u}_1\|_2}.$$

This is why q_{star} is not simply the union of lexical and score directions. It is the lexical direction that overlaps most strongly with the score subspace.

Neuron selection for ablation. Given any unit direction $\mathbf{q} \in \mathbb{R}^d$, we rank neurons by squared loading q_i^2 and select the top-k coordinates,

$$\mathcal{S}_k(\mathbf{q}) = \text{TopK}\left(\{q_i^2\}_{i=1}^d\right).$$

Those coordinates define the neuron set ablated in the causal experiments.

G Full DL19 Neuron Ablation Table

Table 7 reports the full Band A neuron-ablation sweep on DL19, including the smaller $k \in \{16, 32\}$ settings omitted from the main text summary.

Table 7: Full Band A neuron ablation results on TREC DL19. Baseline NDCG@10 = 0.763.

Direction	k	NDCG base	NDCG intv	ΔNDCG	z
q_{lex}	16	0.763	0.762	−0.001	−1.41
q_{lex}	32	0.763	0.761	−0.001	+0.06
q_{lex}	64	0.763	0.762	−0.001	−0.58
q_{lex}	128	0.763	0.760	−0.002	+0.07
q_{score}	16	0.763	0.766	+0.003	+1.49
q_{score}	32	0.763	0.764	+0.001	+0.24
q_{score}	64	0.763	0.761	−0.002	−0.23
q_{score}	128	0.763	0.753	−0.010	−0.36
q_{star}	16	0.763	0.714	−0.049	−30.6
q_{star}	32	0.763	0.706	−0.057	−4.24
q_{star}	64	0.763	0.711	−0.052	−18.6
q_{star}	128	0.763	0.710	−0.053	−12.1

H Per-Query Consistency

Table 8 summarizes how often each intervention worsens, improves, or leaves query-level NDCG unchanged on DL19.

Table 8: Per-query sign summary on DL19.

Direction	k	Worsen	Improve	Unchanged	Median ΔNDCG
q_{lex}	64	7	18	18	0.000
q_{lex}	128	10	17	16	0.000
q_{score}	64	7	15	21	0.000
q_{score}	128	19	13	11	0.000
q_{star}	64	31	8	4	−0.0318
q_{star}	128	30	8	5	−0.0318

I Layer-1 Localization and Control Summary

Table 9 collects the Layer 1 localization results together with the same-layer and exclusion controls referenced in the main text.

Table 9: DL19 localization and control summary. "A" = full Band A (Layers 0 to 15); "L1" = Layer 1 only; "L0 + L2 to L15" = Band A excluding Layer 1. Baseline NDCG@10 = 0.763. All rows from the same control run ($n = 20$ random neuron controls, seed 42).

Direction	Spec	k	ΔNDCG	ΔMAP	z
q_{star}	Band A	64	−0.052	−0.060	−24.0
q_{star}	Layer 1 only	64	−0.049	−0.056	−30.7
q_{star}	L0 + L2 to L15	64	+0.001	−0.001	+0.25
q_{lex}	Layer 1 only	64	≈ 0	+0.001	−0.26
q_{score}	Layer 1 only	64	≈ 0	≈ 0	−0.70
q_{star}	Band A	128	−0.053	−0.055	−8.77
q_{star}	Layer 1 only	128	−0.045	−0.048	−4.00
q_{star}	L0 + L2 to L15	128	−0.004	−0.002	−0.65
q_{lex}	Layer 1 only	128	+0.001	+0.000	+0.53
q_{score}	Layer 1 only	128	≈ 0	−0.002	−0.37

J DL20 Replication Summary

Table 10 reports the held-out DL20 replication results for the same intervention directions used in the main DL19 analysis.

Table 10: DL20 replication summary.

Direction	k	ΔNDCG	ΔMAP	z
q_{lex}	128	+0.004	−0.006	+0.52
q_{score}	128	−0.007	−0.014	−0.93
q_{star}	64	−0.059	−0.117	−18.6
q_{star}	128	−0.059	−0.115	−7.70

K Supplementary Structural Analyses

K.1 Direction Geometry: Semantic Orthogonality at Layer 1

Semantic PLS and vocabulary projection. PLS from Layer 1 activations to gold qrel grades yields only weak alignment with human relevance at this layer ($r = +0.10$). The strongest semantic direction is near-orthogonal to q_{star}, q_{score}, and the score

PCs, with $0/64$ top-neuron overlap with q_{star}. Vocabulary projection through the LM head is also weak and more consistent with template- or format-associated structure than with a clean semantic relevance axis.

K.2 Large-Scale Projection Analysis

If q_{star} were simply approximating BM25 or CQTR, we would expect high correlation between direction activation strength (|proj|) and measured feature values. Over 200,000 MSMARCO dev pairs, every tested correlation is small ($|r| < 0.01$) for BM25, CQTR, mean TF×IDF, stream length, retrieval score, and three PRF features. This does not prove that q_{star} is orthogonal to all classical IR structure. It does show that no single measured feature provides a strong one-dimensional summary of the aligned direction.

K.3 Targeted Diagnostic Benchmark

To better understand what kinds of examples are associated with high or low q_{star} activation, we build a small targeted benchmark from the same projection table used in the large 200K MSMARCO analysis. The goal is not to create a new evaluation benchmark, but to inspect a smaller set of cases more directly. The resulting set contains 320 examples spread across eight buckets. Its labels come from proxy retrieval signals rather than full gold judgments, so we use it only as a diagnostic tool.

At first, this targeted set seemed to support a fairly clean story. Examples in the high-activation proxy-negative buckets appeared to look like lexical traps or topically related but ultimately wrong passages. Examples in the low-activation proxy-positive buckets appeared to look more like indirect or low-overlap semantic matches. However, a later manual audit weakens that story substantially. Across the three proxy-negative buckets, only 4 of 24 audited examples looked truly negative on manual inspection, while 20 of 24 appeared actually relevant.

The targeted follow-up therefore supports only a narrower conclusion than the raw proxy labels initially suggested. q_{star} clearly suppresses easy negatives and often raises scores for answer-like passages. What the benchmark does *not* support is a strong claim that high q_{star} activation cleanly identifies false positives or lexical traps in general. The current diagnostic set is simply too noisy at the proxy-label level for that stronger interpretation.

Table 11: Targeted benchmark follow-up with direct Layer 1 q_{star} ablation. Positive score drop means the q_{star} direction normally raises the example's score; negative score drop means it normally suppresses it.

Bucket	**Mean |proj|**	**Mean drop** k=64	**Mean drop** k=128
fp_lexical_trap	1.656	+0.962	+0.786
fp_constraint_mismatch	0.286	+1.053	+0.773
fp_topical_semantic	0.089	+0.512	+0.292
mp_low_overlap_semantic	0.078	+1.356	+1.220
mp_indirect_contextual	0.078	+1.006	+0.812
mp_definition_alias	0.078	+0.399	+0.304
tp_easy_lexical	0.091	+1.146	+0.912
tn_easy_negative	0.074	-2.219	-2.187