Vector Scores as Likelihood Ratios: Index-Derived Bayesian Calibration for Hybrid Search

Jaepil Jeong
Cognica, Inc.
Email: jaepil@cognica.io
Date: March 23, 2026

"The theory of probabilities is at bottom nothing but common sense reduced to calculus."
— Pierre-Simon Laplace, *Théorie analytique des probabilités*, 1812

Abstract

Vector similarity scores — cosine similarity, inner product, Euclidean distance — are not probabilities. A cosine similarity of 0.85 does not mean an 85% chance of relevance, yet hybrid search systems routinely combine such scores with lexical signals through ad-hoc normalization (min-max, arctangent) or rank-based fusion (RRF) that discards score magnitude information entirely.

We present a Bayesian calibration framework that transforms vector similarity scores into calibrated relevance probabilities by exploiting the distributional statistics already computed during approximate nearest neighbor (ANN) index construction and search. Our approach is grounded in a likelihood ratio formulation: the calibrated probability is determined by the ratio of a *local* distance density (how likely this distance is among relevant documents) to a *global* background density (how likely this distance is by chance in the corpus), combined with an independent prior.

We address the fundamental circularity problem — estimating the local density requires knowing which documents are relevant — through cross-modal conditional independence: any relevance signal conditionally independent of vector distance given true relevance (e.g., lexical matching, alternative embedding models, or index-derived density priors) provides importance weights that break the self-referential loop. We develop two estimation procedures: a nonparametric weighted kernel density estimator for the local distribution, and a parametric Gaussian mixture model with EM optimization initialized by external relevance priors. For pure vector environments where no external signal is available, we present fallback strategies based on distance distribution gap detection and index-derived density priors.

Both estimators derive their statistics from ANN index structures — IVF cell populations and intra-cluster distances, HNSW edge distances and search trajectories — at negligible additional cost. The resulting calibrated vector scores integrate seamlessly with other calibrated signals through additive log-odds fusion (Jeong, 2026a; 2026b), yielding a unified hybrid search framework where every signal contributes independently calibrated Bayesian evidence:

$$\text{logit}\, P(R \mid d_{\text{vec}}, s_{\text{bm25}}) = \underbrace{\log \frac{\hat{f}_R(d)}{f_G(d)}}_{\substack{\text{calibrated} \\ \text{vector evidence}}} + \underbrace{\alpha(s_{\text{bm25}} - \beta)}_{\substack{\text{calibrated} \\ \text{lexical evidence}}} + \underbrace{\text{logit}\, P_{\text{base}}}_{\substack{\text{corpus} \\ \text{prior}}} \tag{1}$$

This completes the probabilistic unification of sparse and dense retrieval: both paradigms are calibrated through the same Bayesian likelihood ratio structure, each drawing on the statistics of its native index, and both contribute as additive evidence in the log-odds space where Bayesian updates are naturally linear. Experimental evaluation on 5 BEIR benchmark datasets using UQA — a multi-paradigm database engine with native IVF implementation — demonstrates: (1) improved calibration over ad-hoc baselines (global sigmoid ECE 0.009 vs arctangent 0.186 on ArguAna); (2) that cross-modal BM25 weights outperform structurally independent density priors despite violating conditional independence (NDCG@10 26.70 vs 20.04 averaged over 5 datasets), validating the pragmatic use of Assumption 4.2.1; and (3) competitive NDCG@10 (41.11 with VPT-calibrated attention fusion) compared to RRF (40.49) and convex combination (41.15) baselines.

---

1. Introduction

1.1 The Vector Score Interpretation Problem

Dense retrieval systems encode queries and documents as vectors in a learned embedding space and rank documents by similarity — typically cosine similarity, inner product, or Euclidean distance (Karpukhin et al., 2020; Khattab & Zaharia, 2020). These similarity scores are routinely treated as relevance indicators, yet they lack a principled probabilistic interpretation.

Problem 1.1.1 (Vector Score Miscalibration). Vector similarity scores suffer from the following interpretability limitations:

1. **Not Probabilities**: A cosine similarity of $s \in [-1, 1]$ is a geometric quantity (the cosine of the angle between two vectors), not a probability of relevance.
2. **Distribution Dependence**: The distribution of similarity scores varies with the embedding model, corpus, and query distribution. A score of 0.7 may be highly discriminative in one corpus and uninformative in another.
3. **Local Density Variation**: The same similarity score carries different information in dense versus sparse regions of the embedding space. In a dense cluster, a nearby document may be unremarkable; in a sparse region, the same distance implies strong relevance.
4. **Scale Incompatibility**: Direct combination with calibrated lexical scores (e.g., Bayesian BM25 probabilities) is unprincipled without a shared probabilistic semantics.

Remark 1.1.2 (The [0, 1] Illusion). Rescaling vector scores to $[0, 1]$ — for instance via $(1 + \cos\theta)/2$ — creates the appearance of probabilities without their substance. A value of 0.8 after linear rescaling does not mean "80% likely relevant." The rescaled score preserves the ranking order of the original cosine similarity but does not calibrate against the true distribution of relevant and non-relevant documents.

1.2 Existing Approaches and Their Limitations

Definition 1.2.1 (Ad-Hoc Normalization). Common score normalization techniques include:

- **Min-max normalization**: $p = \frac{s - s_{\min}}{s_{\max} - s_{\min}}$
- **Arctangent normalization**: $p = \frac{2}{\pi}\arctan(\alpha \cdot s)$
- **Linear rescaling**: $p = \frac{1+s}{2}$ for $s \in [-1, 1]$

Theorem 1.2.2 (Normalization Inadequacy). All query-independent normalization functions fail to account for the local density structure of the embedding space. Formally, for any fixed monotonic transformation $g : \mathbb{R} \to [0, 1]$:

$$g(s_1) = g(s_2) \implies s_1 = s_2 \tag{2}$$

but the true relevance probabilities may differ:

$$P(R = 1 \mid s_1, \text{dense region}) \neq P(R = 1 \mid s_2, \text{sparse region}) \tag{3}$$

even when $s_1 = s_2$.

Proof. A fixed transformation depends only on the score s and is blind to the local density of the embedding space around the query. Two documents equidistant from a query carry different relevance information depending on whether they are in a region with many nearby documents (dense) or few (sparse). No query-independent function can capture this distinction. □

1.3 Structural Parallel with Lexical Retrieval

Bayesian BM25 (Jeong, 2026a) calibrates lexical scores using statistics from the inverted index: document frequency, term frequency, and average document length — all computed at index time. The IDF component is itself a log likelihood ratio:

$$\text{IDF}(t) = \log \frac{N - \text{df}(t) + 0.5}{\text{df}(t) + 0.5} \tag{4}$$

This is the log ratio of the probability of *not* containing term t to the probability of containing it — a density ratio over the term occurrence distribution.

Observation 1.3.1 (Index Statistics Duality). The inverted index provides corpus statistics (df, tf, avgdl) that enable BM25 calibration. ANN indexes (IVF, HNSW) similarly compute and store distributional statistics during construction and search. If lexical calibration exploits inverted index statistics, vector calibration should exploit ANN index statistics. The mathematical structure — a likelihood ratio over corpus distributions — is identical in both cases.

1.4 Our Contribution

We present a framework for calibrating vector similarity scores into relevance probabilities that:

1. **Likelihood Ratio Foundation** (Section 3): Formulates calibration as the ratio of local (relevant) to global (background) distance densities, grounded in Bayes' theorem.
2. **Circularity Resolution** (Section 4): Breaks the self-referential loop in unsupervised local density estimation through cross-modal conditional independence with any external relevance signal.
3. **Nonparametric Estimation** (Section 4): Provides a weighted kernel density estimator for the local distribution using external importance weights.
4. **Pure Vector Fallback** (Section 4.6): Presents estimation strategies — distance gap detection, index density priors, and multi-model cross-calibration — for environments without external signals.
5. **Parametric Estimation** (Section 5): Develops a Gaussian mixture model with EM optimization, using external priors for informed initialization.
6. **Index-Aware Statistics** (Section 6): Extracts the required distributional information from IVF and HNSW index structures at negligible additional cost.

7. **Unified Hybrid Fusion** (Section 7): Integrates calibrated vector scores with other calibrated signals through additive log-odds, completing the probabilistic unification of sparse and dense retrieval.

---

2. Mathematical Preliminaries

2.1 Notation and Conventions

We adopt the notation established in Jeong (2026a, 2026b). Let:

Symbol	Definition
d	Observed distance (or transformed similarity) between query and document vectors
R	Binary relevance variable, $R \in \{0, 1\}$
$f_R(d)$	Probability density of distance d among relevant documents (local distribution)
$f_G(d)$	Probability density of distance d in the full corpus (global/background distribution)
$\sigma(x)$	Sigmoid function: $\frac{1}{1+\exp(-x)}$
$\text{logit}(p)$	Log-odds function: $\log \frac{p}{1-p}$
$\mathcal{K}_h$	Kernel function with bandwidth h
w_i	Lexical prior weight for document i

Convention 2.1.1 (Distance Orientation). Throughout this paper, d denotes a distance-like quantity where *smaller* values indicate *greater* similarity. For cosine similarity $s \in [-1, 1]$, we use $d = 1 - s \in [0, 2]$. For Euclidean distance, d is used directly. The framework applies to any distance metric; only the distributional forms of f_R and f_G change.

2.2 Bayesian BM25 Recap

Theorem 2.2.1 (Three-Term Posterior Decomposition, Jeong 2026a, Theorem 4.4.2). The Bayesian BM25 posterior decomposes as three additive terms in log-odds space:

$$\text{logit}\, P(R = 1 \mid s, f, \hat{n}) = \underbrace{\alpha(s - \beta)}_{\text{logit}(L)} + \underbrace{\text{logit}(b_r)}_{\text{base rate}} + \underbrace{\text{logit}(p)}_{\text{document prior}} \tag{5}$$

where $L = \sigma(\alpha(s - \beta))$ is the sigmoid likelihood, b_r is the corpus-level base rate, and $p = P_{\text{prior}}(f, \hat{n})$ is the composite document prior.

Theorem 2.2.2 (Log-Odds Conjunction, Jeong 2026b, Definition 4.3.1). For n calibrated probability signals, the log-odds conjunction produces the final posterior:

$$P_{\text{final}} = \sigma\left(\frac{1}{n^{1-\alpha}} \sum_{i=1}^{n} \text{logit}(P_i) \right) \tag{6}$$

This resolves the conjunction shrinkage problem (Jeong, 2026b, Theorem 3.2.1) while preserving evidence direction (Theorem 4.5.1, properties iii and iv).

2.3 The Neyman-Pearson Framework

Definition 2.3.1 (Likelihood Ratio). For two hypotheses $H_1 : R = 1$ (relevant) and $H_0 : R = 0$ (non-relevant), the likelihood ratio given observed distance d is:

$$\Lambda(d) = \frac{f_R(d)}{f_G(d)} = \frac{P(d \mid R = 1)}{P(d \mid R = 0)} \tag{7}$$

Theorem 2.3.2 (Neyman-Pearson Sufficiency). The likelihood ratio $\Lambda(d)$ is a sufficient statistic for deciding between H_0 and H_1. No other function of d provides more information about relevance.

Proof. By the Neyman-Pearson lemma, the most powerful test of H_0 versus H_1 at any significance level is the likelihood ratio test. Therefore $\Lambda(d)$ captures all information in d relevant to the relevance decision. □

---

3. Likelihood Ratio Calibration

3.1 The Posterior in Log-Odds Form

Theorem 3.1.1 (Vector Calibration Posterior). Given observed distance d, the posterior probability of relevance is:

$$P(R = 1 \mid d) = \frac{f_R(d) \cdot P(R = 1)}{f_R(d) \cdot P(R = 1) + f_G(d) \cdot P(R = 0)} \tag{8}$$

Transforming to log-odds:

$$\operatorname{logit} P(R = 1 \mid d) = \log \frac{f_R(d)}{f_G(d)} + \operatorname{logit} P(R = 1) \tag{9}$$

Proof. Applying Bayes' theorem and dividing numerator and denominator by $f_G(d) \cdot P(R = 0)$:

$$P(R = 1 \mid d) = \frac{\frac{f_R(d)}{f_G(d)} \cdot \frac{P(R=1)}{P(R=0)}}{\frac{f_R(d)}{f_G(d)} \cdot \frac{P(R=1)}{P(R=0)} + 1} \tag{10}$$

Taking the logit:

$$\begin{aligned} \operatorname{logit} P(R = 1 \mid d) &= \log\left(\frac{f_R(d)}{f_G(d)} \cdot \frac{P(R = 1)}{P(R = 0)} \right) \\ &= \log \frac{f_R(d)}{f_G(d)} + \operatorname{logit} P(R = 1) \quad \square \end{aligned} \tag{11}$$

Remark 3.1.2 (Structural Identity with BM25 Calibration). The log-odds posterior for vector calibration has the same additive structure as Bayesian BM25:

Component	BM25 (Jeong, 2026a)	Vector (this paper)
Evidence	$\alpha(s - \beta)$	$\log \frac{f_R(d)}{f_G(d)}$
Prior	$\text{logit}(p)$	$\text{logit}\, P(R = 1)$
Structure	$\text{logit}(L) = \text{linear}(s)$	$\text{logit}(\Lambda) = \log \text{density ratio}$

Both are likelihood ratios in log-odds space; they differ only in the distributional model over which the ratio is computed. BM25 uses the exponential family parametric model (sigmoid); vector calibration uses the empirical density ratio.

3.2 The Log Density Ratio as Evidence

Definition 3.2.1 (Vector Evidence). The log density ratio constitutes the vector evidence:

$$\text{ev}_{\text{vec}}(d) = \log \frac{f_R(d)}{f_G(d)} \tag{12}$$

Theorem 3.2.2 (Evidence Interpretation). The vector evidence $\text{ev}_{\text{vec}}(d)$ quantifies how much more (or less) likely the observed distance is under the relevant document hypothesis than under the background hypothesis:

- $\text{ev}_{\text{vec}}(d) > 0$: distance d is more consistent with relevance than chance — evidence *for* relevance
- $\text{ev}_{\text{vec}}(d) = 0$: distance d is equally consistent with both hypotheses — no evidence
- $\text{ev}_{\text{vec}}(d) < 0$: distance d is more consistent with chance than relevance — evidence *against* relevance

Proof. Follows directly from the monotonicity of the logarithm:
$\log(f_R/f_G) > 0 \iff f_R(d) > f_G(d)$. □

Remark 3.2.3 (Connection to Pointwise Mutual Information). The vector evidence is the pointwise mutual information between distance d and relevance R:

$$\text{ev}_{\text{vec}}(d) = \log \frac{P(d \mid R = 1)}{P(d)} = \text{PMI}(d; R = 1) \tag{13}$$

when $P(d) \approx P(d \mid R = 0) = f_G(d)$, which holds when the fraction of relevant documents is small. This connects our framework to the information-theoretic foundations of IDF in lexical retrieval, where $\text{IDF}(t) \approx -\log P(t)$ is the self-information of term occurrence.

3.3 Distribution-Free Formulation

Remark 3.3.1 (No Distributional Assumptions). Unlike the parametric sigmoid likelihood in Bayesian BM25 (Jeong, 2026a, Definition 4.1.1), the vector calibration framework makes no parametric assumption about f_R or f_G. The likelihood ratio is defined for any pair of densities over any distance metric. This generality is intentional: embedding spaces are learned, not designed, and their distributional properties are model-dependent. The estimation procedures in Sections 4 and 5 provide concrete methods for computing f_R and f_G from data.

3.4 Concentration of Measure in High Dimensions

Theorem 3.4.1 (Background Distribution Concentration). In high-dimensional embedding spaces ($\dim \gg 1$), the distribution of pairwise distances concentrates around a characteristic value μ_G with standard deviation $\sigma_G = O(1/\sqrt{\dim})$. Formally, for uniformly distributed unit vectors on S^{d-1}, the cosine similarity distribution converges to a Gaussian:

$$f_G(s) \xrightarrow{d\to\infty} \mathcal{N}\left(0, \frac{1}{d}\right) \tag{14}$$

Proof sketch. By the central limit theorem on the components of the inner product of random unit vectors. See Vershynin (2018, Chapter 5) for the precise concentration inequality. □

Corollary 3.4.2 (Background Stability). The narrow concentration of f_G implies that the background density can be estimated with high precision from relatively few samples, and that most of the discriminative power of the likelihood ratio resides in the variation of f_R. This is favorable for our framework: the "easy" distribution (f_G) is stable and cheaply estimated, while the "hard" distribution (f_R) — which encodes the query-dependent relevance structure — requires the more sophisticated estimation techniques of Sections 4 and 5.

---

4. Breaking Circularity: Cross-Modal Estimation of the Local Distribution

4.1 The Circularity Problem

Problem 4.1.1 (Self-Fulfilling Estimation). Estimating $f_R(d)$ — the distance distribution among relevant documents — requires knowing which documents are relevant. In the unsupervised setting (no relevance labels), using the top-K retrieved documents as a proxy for the relevant set introduces a self-fulfilling bias: the estimator assumes its own retrievals are relevant, reinforcing the distance distribution of the retrieval method rather than the true relevance distribution.

Formally, let $\mathcal{D}_K = \{d_1, \ldots, d_K\}$ be the distances of the top-K documents retrieved by vector search. Using these directly as samples from f_R yields:

$$\hat{f}_R^{\text{naive}}(d) = \frac{1}{K}\sum_{i=1}^{K} \mathcal{K}_h(d - d_i) \tag{15}$$

This estimator is biased because $\mathcal{D}_K$ is selected *by the same distance metric* being calibrated, conflating the retrieval criterion with the relevance criterion.

4.2 Conditional Independence Assumption

Assumption 4.2.1 (Cross-Modal Conditional Independence). Given true relevance R, the vector distance D and an external relevance signal W are conditionally independent:

$$P(D, W \mid R) = P(D \mid R) \cdot P(W \mid R) \tag{16}$$

The external signal W may be any relevance indicator that is not derived from the same vector distance being calibrated. Concrete instances include:

- **Lexical matching** (BM25): captures exact term overlap, weighted by corpus statistics (IDF, document length)
- **Alternative embedding model**: a second encoder producing similarity scores from a different learned representation

- **Metadata signals**: document recency, source authority, click-through rates, or other features independent of the embedding geometry

Justification 4.2.2. The conditional independence assumption requires that, given true relevance, the external signal W provides no additional information about the vector distance D. This is plausible when W captures a fundamentally different aspect of relevance than vector proximity. For the canonical case of lexical matching (BM25), vector similarity captures semantic proximity — sensitive to paraphrase, synonymy, and contextual meaning — while BM25 captures exact term overlap. A document may be semantically similar (high vector score) but use different vocabulary (low BM25), or vice versa. Given that the document is truly relevant, the residual correlation between vector distance and lexical score is expected to be small. This is the same structural assumption underlying the naive Bayes classifier, which achieves strong practical performance despite violations of strict independence (Domingos & Pazzani, 1997).

Remark 4.2.3 (Practical Robustness). Even when the conditional independence assumption is violated, the weighted KDE estimator (Definition 4.3.1) degrades gracefully. Residual correlation between D and W given R causes the importance weights to be slightly miscalibrated, but the *direction* of the weighting — upweighting externally relevant documents and downweighting externally irrelevant ones — remains correct as long as the correlation between W and R is non-trivial.

4.3 Weighted Kernel Density Estimation

Definition 4.3.1 (Importance-Weighted KDE). Let $\mathcal{D}_K = \{d_1, \ldots, d_K\}$ be the distances of the top-K documents retrieved by vector search, and let $\mathcal{W}_K = \{w_1, \ldots, w_K\}$ ($w_i \in [0, 1]$) be the corresponding external relevance weights derived from a signal satisfying Assumption 4.2.1 (e.g., Bayesian BM25 probabilities, alternative embedding scores, or index-derived density priors). The importance-weighted kernel density estimate of f_R is:

$$\hat{f}_R(d) = \frac{1}{\sum_{i=1}^{K} w_i} \sum_{i=1}^{K} w_i \, \mathcal{K}_h(d - d_i) \tag{17}$$

where $\mathcal{K}_h$ is a kernel function with bandwidth h.

Theorem 4.3.2 (Circularity Breaking). Under Assumption 4.2.1 (cross-modal conditional independence), the weighted KDE $\hat{f}_R(d)$ is a consistent estimator of the true local distribution $f_R(d)$, where the external weights w_i serve as importance sampling weights that correct for the selection bias of vector retrieval.

Proof. The vector retrieval process selects documents with probability proportional to their vector proximity, introducing a sampling bias relative to the true relevant document distribution. Under conditional independence, the external weight $w_i = P(R = 1 \mid W_i)$ is an unbiased estimator of the relevance probability that is independent of the selection mechanism (which depends on D). Therefore, weighting by w_i performs importance sampling that corrects the selection bias:

$$\begin{aligned} \hat{f}_R(d) &= \frac{\sum_i P(R = 1 \mid W_i) \, \mathcal{K}_h(d - d_i)}{\sum_i P(R = 1 \mid W_i)} \\ &\xrightarrow{K \to \infty} \frac{\mathbb{E}[P(R \mid W) \, \mathcal{K}_h(d - D) \mid D \in \text{top-}K]}{\mathbb{E}[P(R \mid W) \mid D \in \text{top-}K]} \end{aligned} \tag{18}$$

Under conditional independence, $\mathbb{E}[P(R \mid W) \mid D = d_i, R = 1] = \mathbb{E}[P(R \mid W) \mid R = 1]$ is a constant, and the estimator converges to $f_R(d)$ as $K \to \infty$ and $h \to 0$ at appropriate rates. □

Remark 4.3.3 (Intuitive Interpretation). The weighted KDE performs a soft partition of the retrieved documents: those with high external relevance ($w_i \approx 1$) contribute fully to the local density estimate, while those with low external relevance ($w_i \approx 0$) — which are close in vector space but likely false positives — contribute minimally. This suppresses the density contribution of semantically similar but irrelevant documents (e.g., documents in the same topical cluster but not answering the query).

4.4 Bandwidth Selection

Definition 4.4.1 (Silverman's Rule for Weighted KDE). The bandwidth is estimated using the weighted analog of Silverman's rule of thumb:

$$h = 1.06 \cdot \hat{\sigma}_w \cdot K_{\text{eff}}^{-1/5} \tag{19}$$

where $\hat{\sigma}_w$ is the weighted standard deviation of the distances and $K_{\text{eff}} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}$ is the effective sample size accounting for weight variation.

Remark 4.4.2 (Over-Smoothing Risk). Silverman's rule is optimal when the underlying distribution is approximately Gaussian. In high-dimensional embedding spaces, the relevant document distance distribution f_R tends to be sharply peaked — concentrated in a narrow band near the query — due to the concentration of measure (Theorem 3.4.1). Applying Silverman's rule directly may over-smooth this peak, diluting the likelihood ratio f_R / f_G and degrading calibration quality. We recommend a bandwidth ablation study evaluating $h_{\text{eff}} = c \cdot h$ for scaling factors $c \in \{0.2, 0.5, 1.0, 2.0\}$ to identify the optimal smoothing level for the target embedding space. Alternatively, likelihood cross-validation — selecting h to maximize the leave-one-out log-likelihood of the weighted KDE — provides an adaptive, data-driven bandwidth without the Gaussian assumption.

Definition 4.5.1 (Background Density Estimation). The global distribution $f_G(d)$ is estimated from a random sample of query-document distances at index construction time:

$$\hat{f}_G(d) = \frac{1}{M} \sum_{j=1}^{M} \mathcal{K}_{h'}(d - d_j^{\text{rand}}) \tag{20}$$

where $\{d_1^{\text{rand}}, \ldots, d_M^{\text{rand}}\}$ are distances from random queries to their nearest documents.

Remark 4.5.2 (One-Time Computation). The background distribution depends only on the corpus geometry and the embedding model, not on any specific query. It is estimated once during index construction and reused for all queries. By Theorem 3.4.1, the concentration of measure ensures that f_G can be estimated with high precision from relatively few samples ($M \sim 10^3$ typically suffices).

4.6 Estimation Without External Signals

When no external relevance signal is available — for instance, in pure vector search over non-textual data (images, audio, molecular embeddings) — the cross-modal weighting strategy of Section 4.3 cannot be applied. We present three fallback strategies that derive calibration entirely from the vector index and distance distribution.

Strategy 4.6.1 (Distance Gap Detection). The top-K distance distribution often exhibits a natural gap between a cluster of nearby relevant documents and the more distant background. Define the sorted distances $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(K)}$ and the gap sequence $\Delta_i = d_{(i+1)} - d_{(i)}$. The largest gap $\Delta_{i^*} = \max_i \Delta_i$ provides a natural partition: documents with $d \leq d_{(i^*)}$ are treated as the relevant component, and those with $d > d_{(i^*)}$ as background. The GMM (Section 5) can be initialized from this partition rather than from external weights.

Strategy 4.6.2 (Index Density Prior). IVF cell populations provide a density-based relevance prior without requiring any external signal. For a query that probes cell c_j with population n_j:

$$w_i^{\text{density}} = \sigma\left(\gamma \cdot \left(\frac{N/C}{n_j} - 1\right)\right) \tag{21}$$

where N/C is the average cell population and γ is a sensitivity parameter. This assigns higher prior weights to documents in sparse cells (where vector proximity is more discriminative) and lower weights to documents in dense cells (where proximity is less informative). The density prior is conditionally independent of vector distance given relevance — cell assignment is determined at index time, while query-document distance is determined at search time — satisfying Assumption 4.2.1.

Strategy 4.6.3 (Multi-Model Cross-Calibration). When two or more embedding models are available, their similarity scores can serve as mutual external signals. Let $D^{(1)}$ and $D^{(2)}$ be distances from two independently trained encoders. Under the assumption that the encoders capture different aspects of semantic similarity:

$$P(D^{(1)}, D^{(2)} \mid R) \approx P(D^{(1)} \mid R) \cdot P(D^{(2)} \mid R) \tag{22}$$

The score from model 2 serves as w_i for calibrating model 1, and vice versa. This is analogous to co-training (Blum & Mitchell, 1998) in semi-supervised learning.

Remark 4.6.4 (Degradation Hierarchy). The estimation quality degrades gracefully across the strategies: cross-modal weighting (Section 4.3) > index density prior (Strategy 4.6.2) > distance gap detection (Strategy 4.6.1) > naive unweighted KDE (Problem 4.1.1). Even the weakest strategy — distance gap detection — is strictly better than the naive approach because it uses the *structure* of the distance distribution rather than treating all top-K documents as equally relevant.

---

5. Parametric Estimation via Gaussian Mixture Models

5.1 Mixture Model Formulation

For settings where the top-K sample size is small or the nonparametric estimate is noisy, we provide a parametric alternative. We model the distance distribution of the top-K retrieved documents as a two-component Gaussian mixture: one component for relevant documents and one for background (non-relevant) documents that were retrieved due to vector proximity alone.

Definition 5.1.1 (Distance Mixture Model). The distance distribution of the top-K retrieved documents is modeled as:

$$P(d) = \pi \, f_R(d; \theta_R) + (1 - \pi) \, f_G(d; \theta_G) \tag{23}$$

where:

- $\pi \in (0, 1)$ is the mixing coefficient (fraction of relevant documents)
- $f_R(d; \theta_R) = \mathcal{N}(d \mid \mu_R, \sigma_R^2)$ is the relevant document distance distribution
- $f_G(d; \theta_G) = \mathcal{N}(d \mid \mu_G, \sigma_G^2)$ is the background distance distribution
- $\theta_R = (\mu_R, \sigma_R)$ and $\theta_G = (\mu_G, \sigma_G)$ are component parameters

Remark 5.1.2 (Boundary Leakage). Distance metrics are non-negative: cosine distance $d = 1 - s \in [0, 2]$, Euclidean distance $d \in [0, \infty)$. The Gaussian density has support on $(-\infty, \infty)$, so when $\mu_R \approx 0$ (relevant documents very close to the query), the left tail of f_R leaks into the physically impossible negative region. This leakage inflates the normalizing constant and deflates the density at valid distances, potentially distorting the likelihood ratio. In practice, the effect is negligible when $\mu_R / \sigma_R > 3$ (less than 0.1% probability mass below zero). When this condition is violated — which may occur for highly discriminative queries — the Gaussian components should be replaced with non-negative distributions whose support matches the distance metric:

- **Gamma distribution**: $f(d; k, \theta) = \frac{d^{k-1} e^{-d/\theta}}{\theta^k \Gamma(k)}$, $d \geq 0$ — natural for L2 distances
- **Log-normal distribution**: $f(d; \mu, \sigma) = \frac{1}{d\sigma\sqrt{2\pi}} e^{-(\ln d - \mu)^2 / 2\sigma^2}$, $d > 0$ — accommodates right-skewed distance distributions
- **Truncated Gaussian**: $\mathcal{N}(d \mid \mu, \sigma^2) \cdot \mathbf{1}_{[0,\infty)}(d) / Z$ — minimal modification preserving EM tractability

The EM algorithm (Algorithm 5.3.1) applies to all three alternatives with straightforward M-step modifications. The choice between them is an empirical question determined by the shape of the observed distance distribution.

Problem 5.2.1 (Local Optima). The EM algorithm for GMM estimation is sensitive to initialization. In the two-component mixture of relevant and background documents, naive uniform initialization ($\gamma_i^{(0)} = 0.5$ for all i) frequently converges to a solution that separates by distance quantile rather than by relevance, failing to identify the relevant component.

Definition 5.2.2 (Informed Initialization). Let $w_i = P(R = 1 \mid W_i)$ be the relevance probability derived from an external signal W satisfying Assumption 4.2.1. Initialize the E-step responsibilities as:

$$\gamma_i^{(0)} = w_i \tag{24}$$

Theorem 5.2.3 (Convergence Quality). Under Assumption 4.2.1 (cross-modal conditional independence), the informed initialization $\gamma_i^{(0)} = w_i$ provides a strictly better starting point for EM than uniform initialization in the following sense: the initial parameter estimates $\theta_R^{(0)}, \theta_G^{(0)}$ computed from the informed responsibilities have strictly higher likelihood than those from uniform initialization, whenever the external signal has non-trivial correlation with true relevance.

Proof sketch. The initial M-step computes:

$$\mu_R^{(0)} = \frac{\sum_i w_i \, d_i}{\sum_i w_i}, \quad \mu_G^{(0)} = \frac{\sum_i (1 - w_i) \, d_i}{\sum_i (1 - w_i)} \tag{25}$$

When the external weights correlate with relevance, the weighted means separate the relevant component (smaller distances) from the background component (larger distances) more effectively than the unweighted means. The initial log-likelihood $\ell(\theta^{(0)}_{\text{informed}}) > \ell(\theta^{(0)}_{\text{uniform}})$ follows from the improved component separation. Since EM monotonically increases the log-

likelihood, the converged solution inherits this advantage. □

5.3 The EM Algorithm

Algorithm 5.3.1 (EM for Distance Mixture).

> **Input:** Distances $\mathcal{D}_K = \{d_1, \ldots, d_K\}$, external relevance weights $\mathcal{W}_K = \{w_1, \ldots, w_K\}$, background parameters $\theta_G = (\mu_G, \sigma_G)$ from index statistics
>
> **Output:** Estimated f_R parameters $\theta_R = (\mu_R, \sigma_R)$, mixing coefficient π
>
> 1. **Initialize:** $\gamma_i^{(0)} \leftarrow w_i$ for all i; compute $\theta_R^{(0)}$ from weighted M-step
> 2. **Repeat** until convergence:
>
>   **E-step:**
> $\gamma_i^{(t+1)} = \frac{\pi^{(t)}\,\mathcal{N}(d_i;\,\theta_R^{(t)})}{\pi^{(t)}\,\mathcal{N}(d_i;\,\theta_R^{(t)}) + (1-\pi^{(t)})\,\mathcal{N}(d_i;\,\theta_G)}$
>
>   **M-step:**
> $\pi^{(t+1)} = \frac{1}{K}\sum_i \gamma_i^{(t+1)}$
> $\mu_R^{(t+1)} = \frac{\sum_i \gamma_i^{(t+1)}\, d_i}{\sum_i \gamma_i^{(t+1)}}$
> $\sigma_R^{(t+1)} = \sqrt{\frac{\sum_i \gamma_i^{(t+1)} (d_i - \mu_R^{(t+1)})^2}{\sum_i \gamma_i^{(t+1)}}}$
>
> 3. **Return** $\theta_R = (\mu_R, \sigma_R)$, π

Remark 5.3.2 (Fixed Background Component). The background parameters θ_G are fixed to the index-time estimates (Section 6) and not updated during EM. This halves the parameter space and prevents the background component from absorbing relevant documents.

5.4 Relationship Between Sections 4 and 5

Remark 5.4.1 (Nonparametric vs. Parametric). Sections 4 and 5 present two estimation strategies for the same quantity $f_R(d)$:

Property	Weighted KDE (Section 4)	GMM-EM (Section 5)
Distributional assumption	None (nonparametric)	Gaussian (parametric)
Minimum sample size	$K \geq 50$ (rule of thumb)	$K \geq 10$
Sensitivity to outliers	Moderate (bandwidth-dependent)	Low (Gaussian smoothing)
Computation cost	$O(K)$ per query point	$O(K \cdot T_{\text{EM}})$ per query
Flexibility	High (arbitrary shapes)	Low (unimodal Gaussian per component)

Both estimators use the same external relevance weights w_i to break circularity; they differ in the functional form imposed on the estimate. The nonparametric estimator is preferred when K is large and the true distribution may be non-Gaussian; the parametric estimator is preferred when K is small and robustness to sampling noise is critical. When no external signal is available, the

fallback strategies of Section 4.6 provide alternative weight sources.

---

6. Index-Aware Statistics Extraction

6.1 The Zero Additional Cost Principle

Observation 6.1.1 (Free Statistics). Both IVF and HNSW index structures compute distributional statistics during construction and search that are directly useful for calibration. Extracting these statistics incurs negligible additional computational cost because the underlying quantities are already computed and typically discarded.

6.2 IVF Index Statistics

Definition 6.2.1 (IVF Structure). An IVF (Inverted File) index partitions the corpus into C cells via k-means clustering. Each cell c_j ($j = 1, \ldots, C$) contains vectors assigned to centroid μ_j, with cell population $n_j = |c_j|$.

Theorem 6.2.2 (IVF Statistics for Calibration). The following statistics are computed during IVF construction and search, and directly estimate the distributions needed for calibration:

(a) Global Background Distribution f_G:

- **Inter-centroid distances**: The pairwise distances between centroids $\{\|\mu_i - \mu_j\|\}_{i \neq j}$ provide the macro-scale structure of the embedding space.
- **Intra-cell distance distributions**: For each cell c_j, the distances $\{d(v, \mu_j) : v \in c_j\}$ characterize the local spread. The aggregate across all cells estimates f_G.

(b) Local Distribution Inputs $\{d_i\}$ for f_R Estimation:

- **Probed cell distances**: During search, the IVF probes n_{probe} cells nearest to the query. The distances within these cells are already computed for top-K selection.

(c) Local Density Prior:

- **Cell population n_j**: The number of vectors in the probed cell provides a density estimate. A match in a densely populated cell (large n_j) is less informative than a match in a sparse cell (small n_j), analogous to inverse document frequency in lexical retrieval.

Definition 6.2.3 (Cell-Aware Base Rate). The local base rate within cell c_j can be adjusted by cell density:

$$P_{\text{base}}^{(j)} = P_{\text{base}} \cdot \frac{N/C}{n_j} \tag{26}$$

This upweights matches in sparse cells (where proximity is more meaningful) and downweights matches in dense cells (where proximity is less discriminative).

6.3 HNSW Index Statistics

Definition 6.3.1 (HNSW Structure). An HNSW (Hierarchical Navigable Small World) graph index maintains a multi-layer proximity graph where each layer ℓ connects vectors to their approximate nearest neighbors, with layer 0 being the densest.

Theorem 6.3.2 (HNSW Statistics for Calibration). The following statistics are available from HNSW construction and search:

(a) Global Background Distribution f_G:

- **Edge distance distribution**: The distances along all edges in layer 0, $\{d(v_i, v_j) : (v_i, v_j) \in E_0\}$, provide a biased but informative sample of pairwise distances in the corpus. This distribution is computed at build time and stored implicitly in the graph structure.

(b) Local Distribution Inputs $\{d_i\}$ for f_R Estimation:

- **Search trajectory distances**: During greedy search, HNSW computes distances to all candidate nodes visited. The set of visited-node distances — currently used only for top-K selection and then discarded — constitutes a natural sample from the query's local neighborhood.

(c) Local Density Proxy:

- **Search radius**: The distance to the K-th nearest neighbor found during search provides a direct measure of local density: small radius implies dense region, large radius implies sparse region.
- **Visited node count**: The number of distance computations required during search correlates inversely with local density (dense regions require fewer hops).

6.4 Index Statistics Mapping

Table 6.4.1 (Unified Statistics Mapping).

Required Statistic	IVF Source	HNSW Source	Cost
$f_G(d)$ — global background	Intra-cell distance aggregate	Layer-0 edge distance distribution	Build-time only
$\{d_i\}$ — local distances	Probed cell distances	Search trajectory distances	Already computed
Local density proxy	Cell population n_j	Search radius / visited count	Already computed
Query-region prior	Query-to-centroid distance	Entry-point distance	Already computed

Remark 6.4.2 (Universality). The calibration framework (Theorem 3.1.1) is expressed in terms of abstract densities f_R and f_G, independent of the index structure. Table 6.4.1 shows that both major ANN index families provide the required statistics as search by-products. The framework is therefore index-agnostic in its mathematical formulation and index-aware only in its estimation procedure.

---

7. Unified Hybrid Search Fusion

7.1 The Complete Log-Odds Decomposition

Theorem 7.1.1 (Unified Posterior). Given a query q and document d with BM25 score s_{bm25}, vector distance d_{vec}, and corpus prior P_{base}, the posterior probability of relevance under cross-modal conditional independence is:

$$\operatorname{logit} P(R = 1 \mid s_{\text{bm25}}, d_{\text{vec}}) = \underbrace{\log \frac{\hat{f}_R(d_{\text{vec}})}{f_G(d_{\text{vec}})}}_{\substack{\text{calibrated} \\ \text{vector evidence}}} + \underbrace{\alpha(s_{\text{bm25}} - \beta)}_{\substack{\text{calibrated} \\ \text{lexical evidence}}} + \underbrace{\operatorname{logit} P_{\text{base}}}_{\substack{\text{corpus} \\ \text{prior}}} \tag{27}$$

Proof. Under conditional independence of D and S_{bm25} given R:

$$P(R \mid D, S) = \frac{P(D \mid R)\, P(S \mid R)\, P(R)}{P(D, S)} \tag{28}$$

Taking the logit and using the factorization:

$$\operatorname{logit} P(R \mid D, S) = \log \frac{P(D \mid R = 1)}{P(D \mid R = 0)} + \log \frac{P(S \mid R = 1)}{P(S \mid R = 0)} + \operatorname{logit} P(R) \tag{29}$$

The first term is the vector evidence $\log(f_R / f_G)$. The second term is the BM25 evidence, which under the sigmoid likelihood model (Jeong, 2026a, Definition 2.3.2) equals $\alpha(s - \beta)$. The third term is the corpus prior. □

Corollary 7.1.2 (Extensibility). The additive structure admits arbitrary additional signals. For n conditionally independent signals:

$$\operatorname{logit} P(R \mid s_1, \ldots, s_n) = \sum_{i=1}^{n} \underbrace{\log \frac{f_{R,i}(s_i)}{f_{G,i}(s_i)}}_{\text{evidence from signal } i} + \operatorname{logit} P_{\text{base}} \tag{30}$$

Each signal contributes its own likelihood ratio, calibrated using the statistics of its own index or scoring mechanism. New signals (e.g., graph centrality, temporal recency, user behavior) are added by appending their calibrated log-odds term — no re-tuning of existing signals is required.

7.2 Structural Unification of Sparse and Dense Retrieval

Theorem 7.2.1 (Common Mathematical Structure). Both BM25 and vector calibration instantiate the same abstract pattern:

$$\text{evidence}(s) = \log \frac{P(s \mid R = 1)}{P(s \mid R = 0)} \tag{31}$$

For BM25, the log likelihood ratio is parametrized by the sigmoid model: $\alpha(s - \beta)$.

For vector search, the log likelihood ratio is the empirical density ratio: $\log(\hat{f}_R(d) / f_G(d))$.

Proof. Both are instances of Bayes' theorem applied to the same binary relevance variable R, differing only in the distributional model of the observed signal. The log-odds additivity follows from the product rule for conditionally independent evidence. □

Remark 7.2.2 (From Ad-Hoc to Principled). Existing hybrid search methods combine BM25 and vector scores through:

Method	Formula	Weaknesses
Convex combination	$\lambda \cdot s_{\text{bm25}} + (1-\lambda) \cdot s_{\text{vec}}$	Requires tuning λ; incommensurable scales
RRF	$\sum_i \frac{1}{k+\text{rank}_i}$	Discards score magnitudes; arbitrary k
Min-max + linear	Normalize then add	Query-dependent normalization; no probabilistic basis

The log-odds fusion (Theorem 7.1.1) eliminates these weaknesses: both signals are calibrated to a common probabilistic scale, the combination is Bayesian optimal under the conditional independence assumption, and no free parameters require tuning beyond the per-signal calibration parameters.

7.3 Connection to Neural Network Structure

Remark 7.3.1 (Two-Layer Network Revisited). Jeong (2026b, Theorem 5.2.1 and Remark 5.2.3) proved that combining heterogeneously calibrated signals produces a genuine two-layer neural network with logit hidden nonlinearity. In that analysis, the vector signal used a linear calibration $(1+\cos\theta)/2$, making the logit transformation nonlinear:

$$\text{logit}\left(\frac{1+\cos\theta}{2}\right) = \log\frac{1+\cos\theta}{1-\cos\theta} \tag{32}$$

The likelihood ratio calibration of this paper provides a principled replacement: instead of the linear rescaling followed by logit, the vector signal's contribution is the empirical log density ratio $\log(f_R/f_G)$, which is directly a nonlinear function of the distance. The neural network identification remains valid, but with a calibration that is grounded in distributional statistics rather than a geometric rescaling.

---

8. Experimental Design

8.1 Evaluation Framework

Experiments are conducted using UQA (Unified Query Algebra), a multi-paradigm database engine with native IVF implementation in NumPy, enabling direct access to index internals. Evaluation datasets are drawn from the BEIR benchmark suite (Thakur et al., 2021), consistent with the Bayesian BM25 evaluation (Jeong, 2026a, Section 11).

8.2 Baselines

Method	Description
Linear rescaling	$p = (1+\cos\theta)/2$ (Jeong, 2026a, Definition 7.1.2)
Min-max normalization	$p = (s - s_{\min})/(s_{\max} - s_{\min})$ per query
Platt scaling	$p = \sigma(a \cdot s + b)$ with supervised (a, b)
Arctangent normalization	$p = \frac{2}{\pi}\arctan(\alpha \cdot s)$

8.3 Metrics

- **Expected Calibration Error (ECE)**: Measures calibration quality — whether predicted probabilities match empirical frequencies.
- **Log Loss (Negative Log-Likelihood)**: $\mathcal{L} = -\frac{1}{N}\sum_i [y_i \log p_i + (1-y_i)\log(1-p_i)]$. Strictly proper scoring rule that penalizes the probabilistic model directly, without the binning artifacts of ECE.
- **Brier Score**: Joint measure of calibration and refinement.
- **NDCG@10**: Ranking quality metric, ensuring calibration does not degrade retrieval effectiveness.

8.4 Experimental Procedure

1. **Global κ baseline**: Sigmoid calibration $P = \sigma(\kappa(s - \beta))$ with corpus-level κ
2. **IVF cell-aware calibration**: Local f_R estimation using cell statistics (Section 6.2)
3. **Index density prior fallback**: Pure vector calibration using Strategy 4.6.2
4. **Weighted KDE with BM25 weights**: Nonparametric f_R estimation with cross-modal importance weights (Section 4.3)
5. **GMM-EM refinement**: Parametric f_R estimation with informed initialization (Section 5)
6. **Conditional independence penalty**: Direct comparison of stages 3 and 4 — index density prior (structurally independent of vector distance) versus BM25 weights (approximately independent) — to quantify the empirical cost of conditional independence violation against the information gain from lexical signal
7. **Bandwidth ablation**: Scaling factors $c \in \{0.2, 0.5, 1.0, 2.0\}$ applied to Silverman bandwidth (Remark 4.4.2)
8. **Unified log-odds fusion**: Each calibration method combined with Bayesian BM25 via Theorem 7.1.1

Each stage isolates the marginal contribution of the corresponding component.

8.5 Experimental Results

Experiments are evaluated on 5 BEIR datasets (ArguAna, FiQA, NFCorpus, SciDocs, SciFact) using the exact dense backend with all-MiniLM-L6-v2 embeddings. BM25 parameters: $k_1 = 1.2$, $b = 0.75$, Lucene variant with Snowball English stemmer. Retrieval depth $R = 1000$, evaluation depth $k = 10$.

8.5.1 Calibration Baselines (Stage 1)

Table 8.5.1. ECE of vector score calibration baselines. All methods are monotone transforms and produce identical NDCG@10 to raw Dense retrieval (38.32 average). Lower ECE is better.

Method	ArguAna	FiQA	NFCorpus	SciDocs	SciFact
Dense-Kappa (global κ)	**0.009**	**0.021**	0.231	0.210	**0.032**
Dense-Arctan	0.186	0.237	0.463	0.132	0.232
Dense-Platt (supervised)	0.065	0.075	**0.097**	**0.165**	0.074

The unsupervised global sigmoid (κ baseline) achieves the lowest ECE on 3 of 5 datasets, outperforming the supervised Platt scaling baseline. This is consistent with the observation that corpus-level distance statistics (Theorem 3.4.1) provide a strong calibration signal without requiring relevance labels.

8.5.2 Conditional Independence Penalty (Stage 6)

Table 8.5.2. NDCG@10 comparison of CI-compliant vs CI-violating estimation strategies. Both methods are fused with Bayesian BM25 via Theorem 7.1.1.

Method	ArguAna	FiQA	NFCorpus	SciDocs	SciFact	Average
VPT-DensityPrior (CI-compliant)	1.66	17.76	25.90	12.75	42.12	20.04
VPT-BM25Weights (CI-violating)	0.02	**24.38**	**35.61**	**13.53**	**59.95**	**26.70**

Table 8.5.3. ECE comparison of the same methods.

Method	ArguAna	FiQA	NFCorpus	SciDocs	SciFact	Average
VPT-DensityPrior (CI-compliant)	**0.040**	0.204	0.928	0.251	0.272	0.339
VPT-BM25Weights (CI-violating)	0.000	**0.078**	0.853	**0.176**	**0.060**	**0.233**

The BM25-weighted estimator outperforms the density-prior-only estimator on both ranking quality (NDCG@10: 26.70 vs 20.04) and calibration quality (ECE: 0.233 vs 0.339) in aggregate, despite violating the conditional independence assumption (Assumption 4.2.1). The results exhibit a revealing bidirectional pattern:

- **On 4 of 5 datasets** (FiQA, NFCorpus, SciDocs, SciFact), BM25 weights dominate density priors across all metrics. The information gain from cross-modal lexical signal outweighs the bias introduced by residual dependence.
- **On ArguAna**, the relationship reverses: VPT-DensityPrior (1.66 NDCG@10) outperforms VPT-BM25Weights (0.02). ArguAna is a counter-argument retrieval dataset where the relevant document is the argument *against* the query — making BM25 lexical similarity an adversarial relevance signal. The calibration framework correctly propagates these misleading weights into degraded probabilities.

This bidirectional result validates the likelihood ratio calibration framework itself: it faithfully reflects the quality of its input signals. When external weights are informative (BM25 on standard retrieval tasks), the calibration amplifies them into improved ranking; when weights are adversarial, the calibration exposes the harm rather than masking it. The practical implication is that the choice of importance weighting signal (Section 4.3) is a modular design decision — the calibration framework is agnostic to the weight source and propagates signal quality transparently.

8.5.3 Bandwidth Ablation (Stage 7)

Table 8.5.4. NDCG@10 for KDE bandwidth scaling factor c applied to Silverman bandwidth (Remark 4.4.2). All methods use BM25 importance weights and are fused with Bayesian BM25.

c	ArguAna	FiQA	NFCorpus	SciDocs	SciFact	Average
0.2	0.02	**28.01**	35.32	**16.54**	**65.92**	**29.16**
0.5	0.02	27.90	35.16	16.49	65.91	29.10
1.0	0.02	27.95	35.03	16.42	65.33	28.95
2.0	0.02	27.27	**35.36**	15.80	63.33	28.36

All bandwidth variants produce 0.02 NDCG@10 on ArguAna, confirming that KDE estimation with BM25 importance weights inherits the adversarial signal quality observed in Stage 6. On the remaining 4 datasets, narrower bandwidths ($c = 0.2$) produce marginally better ranking quality, consistent with the concentration of f_R in high-dimensional embedding spaces (Theorem 3.4.1): the relevant document distribution is sharply peaked, and standard Silverman bandwidth over-smooths this peak, diluting the likelihood ratio f_R / f_G. The differences are modest (29.16 vs 28.36 overall), suggesting that the density ratio is robust to bandwidth selection within a reasonable range.

8.5.4 Unified Log-Odds Fusion (Stage 8)

Table 8.5.5. NDCG@10 for hybrid fusion methods. VPT methods calibrate dense probabilities via the likelihood ratio framework (Theorem 3.1.1) before fusion with Bayesian BM25. Baselines use ad-hoc normalization.

Method	ArguAna	FiQA	NFCorpus	SciDocs	SciFact	Average
BM25	36.13	25.31	31.82	15.63	68.02	35.38
Dense	36.98	36.87	31.59	21.64	64.51	38.32
Convex ($\lambda = 0.5$)	40.01	37.10	35.60	19.67	73.37	41.15
RRF ($k = 60$)	39.61	36.85	34.43	20.11	71.43	40.49
Bayesian-Vector-Balanced	27.39	33.67	29.50	18.51	66.06	35.03
Bayesian-Vector-Softplus	22.47	34.43	32.15	18.94	68.56	35.31
Bayesian-Vector-Attn	**37.66**	**39.81**	**34.82**	**21.94**	**71.34**	**41.11**
Bayesian-Balanced	37.27	40.58	35.73	21.42	72.47	41.50
Bayesian-Attn-Norm	**37.22**	**40.53**	**35.42**	**21.91**	**73.24**	**41.67**

Bayesian-Vector-Attn — which combines VPT-calibrated dense probabilities with Bayesian BM25 via learned attention weights — achieves 41.11 average NDCG@10, competitive with the convex combination baseline (41.15) and RRF (40.49). When both signals use the full Bayesian calibration pipeline (Bayesian-Attn-Norm), performance reaches 41.67, exceeding all baselines.

The simpler VPT fusion variants (Vector-Balanced, Vector-Softplus) underperform because the additive log-odds combination without learned signal weighting is sensitive to the relative scale of lexical and vector evidence. The attention mechanism (Jeong, 2026b, Section 8) resolves this by learning query-dependent weights that adaptively balance the two calibrated signals.

---

9. Discussion

9.1 The Index as a Statistical Model

A recurring theme of this paper is that index structures — inverted files, IVF, HNSW — are not merely data structures for efficient retrieval; they are implicit statistical models of the corpus. The inverted index encodes term frequency statistics that BM25 exploits for scoring. The IVF index encodes cluster structure and distance distributions that this paper exploits for calibration. The unification is not accidental: both arise from the same Bayesian reasoning applied to different data modalities.

9.2 Relationship to Prior Work

This work extends the Bayesian BM25 calibration framework (Jeong, 2026a) from lexical to dense retrieval. Where Bayesian BM25 exploits inverted index statistics (document frequency, term frequency, average document length) to calibrate BM25 scores via a sigmoid likelihood model, the present paper exploits ANN index statistics (cell populations, edge distances, search trajectories) to calibrate vector scores via a density ratio model. The mathematical structure — a likelihood ratio in log-odds space — is shared; only the distributional model changes.

The log-odds conjunction framework developed in Jeong (2026b) provides the fusion mechanism: calibrated signals from heterogeneous sources combine as additive terms in the natural parameter space of Bernoulli random variables. That work proved that this fusion has the computational structure of a feedforward neural network when signals have heterogeneous calibrations — a result that applies directly to the hybrid fusion of Theorem 7.1.1, where the BM25 signal uses sigmoid calibration and the vector signal uses density ratio calibration.

The algebraic foundation — posting lists as a universal abstraction across relational, text, vector, and graph paradigms (Jeong, 2023; 2024) — provides the compositional substrate over which calibrated scores are evaluated. The present paper does not modify this algebraic layer; it enriches the scoring layer that operates above it.

9.3 Limitations and Future Work

1. **Conditional Independence Violation**: The cross-modal conditional independence assumption (Assumption 4.2.1) is approximate. Embedding models trained on text inevitably encode some lexical information, introducing residual correlation. We quantify this penalty empirically (Section 8.4, stage 6) by comparing the structurally independent index density prior (Strategy 4.6.2) against BM25-weighted estimation (Section 4.3). If BM25 weights yield superior calibration despite violating conditional independence, the information gain from lexical signal dominates the bias from dependence — validating the pragmatic use of the assumption. The theoretical question of optimal correction for known dependence structure remains open.
2. **Query-Dependent f_R**: The local distribution varies per query. For frequent query types, caching estimated distributions may improve efficiency; for rare queries, the estimation may be noisy.
3. **Multimodal f_R**: For ambiguous queries, the relevant document distribution may be multimodal (multiple relevant clusters). The two-component GMM (Section 5) captures one relevant mode; extensions to k-component mixtures may be needed.
4. **Online Adaptation**: The framework currently performs per-query estimation. Accumulating calibration statistics across queries — learning f_R and f_G jointly from query logs — is a natural extension.
5. **HNSW Experimental Validation**: The present paper provides full experimental validation on IVF and a theoretical analysis for HNSW (Section 6.3). Empirical validation on HNSW is

deferred to future work.

6. **Pure Vector Fallback Evaluation**: The fallback strategies of Section 4.6 are presented as theoretically motivated alternatives; their empirical calibration quality relative to cross-modal weighting requires systematic evaluation across non-textual domains (image retrieval, molecular search).

---

10. Conclusion

We have presented a Bayesian framework for calibrating vector similarity scores into relevance probabilities, grounded in the likelihood ratio of local and global distance distributions. The key contributions are:

1. **Likelihood Ratio Formulation**: Vector calibration as $\log(f_R/f_G)$ — structurally identical to IDF in lexical retrieval.
2. **Circularity Resolution**: Cross-modal conditional independence with external importance weights breaks the self-referential loop in unsupervised density estimation; fallback strategies enable calibration even without external signals.
3. **Index-Aware Statistics**: IVF and HNSW indexes provide all required distributional statistics as by-products of construction and search, at negligible additional cost.
4. **Unified Fusion**: Calibrated vector and lexical scores combine as additive log-odds, completing the probabilistic unification of sparse and dense retrieval.

The resulting framework fulfills the promise implicit in the Probability Ranking Principle (Robertson, 1977): every ranking signal — lexical, semantic, structural — contributes calibrated probabilistic evidence, combined through Bayesian inference in the natural parameter space of relevance. The index is not merely an optimization structure; it is a statistical model.

References

1. Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *Proceedings of the 11th Annual Conference on Computational Learning Theory*, 92-100.
2. Cormack, G. V., Clarke, C. L., & Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd International ACM SIGIR Conference*, 758-759.
3. Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3), 103-130.
4. Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8), 1771-1800.
5. Jégou, H., Douze, M., & Schmid, C. (2011). Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1), 117-128.
6. Jeong, J. (2023). A unified mathematical framework for query algebras across heterogeneous data paradigms. *OSF Preprints*. https://doi.org/10.31219/osf.io/f56j2_v2
7. Jeong, J. (2024). Extending the unified mathematical framework to support graph data structures. *OSF Preprints*. https://doi.org/10.31219/osf.io/cgfae_v1
8. Jeong, J. (2026a). Bayesian BM25: A probabilistic framework for hybrid text and vector

search. *Zenodo*. https://doi.org/10.5281/zenodo.18414940

9. Jeong, J. (2026b). From Bayesian inference to neural computation: The analytical emergence of neural network structure from probabilistic relevance estimation. *Zenodo*. https://doi.org/10.5281/zenodo.18512411
10. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535-547.
11. Karpukhin, V., et al. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6769-6781.
12. Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference*, 39-48.
13. Malkov, Y. A., & Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4), 824-836.
14. Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 61-74.
15. Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, 33(4), 294-304.
16. Robertson, S. E., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333-389.
17. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *Proceedings of the 35th Conference on Neural Information Processing Systems*.
18. Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.