A Unified Bayesian Framework for Hybrid Search: Calibration and Log-Odds Fusion of Lexical and Vector Retrieval

Jaepil Jeong
Cognica, Inc.
Email: jaepil@cognica.io
Date: Jun 12, 2026

Abstract

Modern retrieval systems combine heterogeneous relevance signals — lexical scores such as BM25, dense vector similarities, and learned re-rankers — yet these signals live on incompatible scales and carry no shared probabilistic meaning. A BM25 score of 8.4 and a cosine similarity of 0.85 cannot be added, averaged, or thresholded together without resorting to ad-hoc normalization or rank-based fusion that discards magnitude information.

This paper develops a single Bayesian principle that resolves the problem for every signal at once: each raw score is the observable of a binary relevance hypothesis, and the calibrated probability of relevance is the posterior obtained from the score's *log-likelihood ratio* plus an independent prior, evaluated in log-odds space. We instantiate this principle twice. For lexical retrieval, a parametric sigmoid likelihood and a corpus-level base rate make the posterior additive in log-odds and reduce expected calibration error by 68–77% without relevance labels. For dense retrieval, a nonparametric likelihood ratio between a local relevant-distance density and a global background density calibrates vector scores using only the distributional statistics that an approximate-nearest-neighbor index already computes; a cross-modal conditional-independence argument breaks the circularity inherent in estimating the relevant density.

Because each signal's evidence is a log-likelihood ratio on a common Bernoulli log-odds scale, fusion is additive: the evidence terms sum, with the prior added once. We develop a log-odds conjunction that aggregates calibrated evidence while avoiding the shrinkage pathology of naive probabilistic conjunction — exactly a normalized Logarithmic Opinion Pool — together with a query-adaptive weighting that lets the per-signal reliabilities depend on query features. The result is a single additive posterior in log-odds that fuses an arbitrary number of conditionally independent signals — here, sparse and dense — with no re-calibration of the existing signals, and preserves the safe dynamic pruning (WAND, Block-Max WAND) of the underlying BM25 retrieval. On five BEIR datasets the framework is competitive with, and slightly exceeds on aggregate, the strongest tuning-free baselines (convex combination, reciprocal rank fusion) while additionally producing calibrated probabilities (best zero-shot NDCG@10 $+6.28$ over BM25; the base-rate correction reduces expected calibration error by 68–77% without relevance labels).

Keywords: information retrieval, hybrid search, probability calibration, Bayesian inference, BM25, vector search, likelihood ratio, log-odds fusion, Logarithmic Opinion Pool, dynamic pruning.

1. Introduction

1.1 The score interpretation problem

Ranking functions in information retrieval produce *scores*, not *probabilities*. This distinction is usually harmless when a single ranker operates in isolation — only the induced order matters — but it becomes the central obstacle the moment two or more signals must be combined, or a hard relevance decision must be made.

Lexical scores. The BM25 function (Robertson and Zaragoza, 2009) assigns a document d and query $q = \{t_1, \ldots, t_m\}$ the score

$$\mathrm{BM25}(d,q) \;=\; \sum_{t\in q} \mathrm{IDF}(t)\;\cdot\; \frac{f(t,d)\,(k_1+1)}{f(t,d)+k_1\!\left(1-b+b\,\frac{|d|}{\mathrm{avgdl}}\right)},$$

where $f(t,d)$ is the term frequency, $|d|$ the document length, avgdl the average length, and k_1, b free parameters. The score is effective for ranking but resists interpretation: its range is unbounded, its magnitude depends on query length and term specificity, and its inverse document frequencies depend on collection statistics, so a score of 8.4 has no fixed meaning across queries or corpora and cannot be compared to a bounded signal such as a cosine similarity.

Vector scores. Dense retrieval scores a query–document pair by the geometric proximity of their embeddings — cosine similarity, inner product, or a monotone function of Euclidean distance. These too are not probabilities. A cosine similarity of 0.85 does not denote an 85% chance of relevance; the value depends on the embedding model, the corpus density around the query, and the intrinsic dimensionality of the representation. Rescaling cosine similarity to $[0,1]$ by $p = (1 + \cos)/2$ produces a number in the unit interval, but a number in $[0,1]$ is not a calibrated probability — the transform is query-independent and ignores how distances are actually distributed.

In both cases the underlying defect is identical: a raw score is an *uncalibrated observable*, and turning it into a probability of relevance requires a statistical model relating the score to the latent binary event "this document is relevant."

1.2 The multi-signal fusion problem

Hybrid search must combine signals of different provenance and scale. The prevailing methods all sacrifice information or principle:

- **Convex combination**, $w \cdot s_{\text{dense}} + (1-w) \cdot s_{\text{sparse}}$, requires tuning w and presupposes the two scores are commensurable, which they are not.
- **Reciprocal Rank Fusion** (RRF) (Cormack et al., 2009), $\sum_\ell 1/(k + \mathrm{rank}_\ell(d))$, discards score magnitudes entirely, retaining only ranks, and introduces an unjustified constant k. It is also non-commutative with filtering: applying a relevance threshold before versus after fusion yields different result sets.
- **Min–max normalization followed by linear combination** rescales each signal into $[0,1]$ per query, but the rescaling is query-dependent, lacks any probabilistic basis, and is sensitive to outliers.

A naïve additive combination of raw scores fails for a more basic reason.

Proposition 1.1 (signal dominance). *For an additive combination $s = s_1 + s_2$ of two zero-mean signals with standard deviations $\sigma_1 \gg \sigma_2$, the probability that s_2 changes the induced ranking of a random pair tends to zero as $\sigma_1/\sigma_2 \to \infty$.*

The signal with the larger dynamic range silently dominates, regardless of which signal is more informative. Calibrating every signal to a common probabilistic scale removes the dependence on dynamic range and replaces ad-hoc weighting with Bayesian evidence combination.

1.3 Contributions and structure

This paper advances a single thesis: **every retrieval signal can be calibrated to a Bayesian relevance probability through its log-likelihood ratio, and calibrated signals combine additively in log-odds space.** Concretely:

1. **A sigmoid-likelihood calibration of lexical scores** (Section 3) with a corpus-level base rate, yielding an additive log-odds posterior that is provably rank-preserving and compatible with WAND/BMW dynamic pruning, plus an optional document-quality prior as a further additive term.
2. **A likelihood-ratio calibration of vector scores** (Section 4) that uses only the distance statistics an ANN index already maintains and addresses the circularity of relevant-density estimation through a cross-modal conditional-independence argument; we treat it as a pragmatic estimator, since the independence assumption is typically only approximate, and it degrades gracefully when no external signal is available.
3. **A log-odds conjunction for fusion** (Section 5) that avoids the shrinkage pathology of naïve probabilistic conjunction, is exactly a normalized Logarithmic Opinion Pool, admits query-adaptive per-signal weights, and yields a single additive posterior that fuses an *arbitrary number* of conditionally independent signals — sparse, dense, or

otherwise — with no re-calibration of existing signals as they are added.

4. **An empirical evaluation** (Section 6) on five BEIR datasets showing label-free calibration gains, ranking quality competitive with strong tuning-free fusion baselines, and the predicted behavior of each component, with an explicit account of the comparison's limitations.

We begin with the probabilistic machinery common to the entire development.

---

2. Mathematical preliminaries

2.1 The sigmoid and logit functions

Definition 2.1 (sigmoid). The logistic sigmoid $\sigma : \mathbb{R} \to (0, 1)$ is

$$\sigma(x) \;=\; \frac{1}{1 + e^{-x}}.$$

It satisfies $\sigma(-x) = 1 - \sigma(x)$ (point symmetry about $(0, \frac{1}{2})$), $\sigma'(x) = \sigma(x)\big(1 - \sigma(x)\big)$, and $\lim_{x\to-\infty} \sigma(x) = 0$, $\lim_{x\to+\infty} \sigma(x) = 1$.

Definition 2.2 (logit). The logit (log-odds) function $\operatorname{logit} : (0, 1) \to \mathbb{R}$ is the inverse of the sigmoid,

$$\operatorname{logit}(p) \;=\; \log \frac{p}{1 - p}, \qquad \sigma\big(\operatorname{logit}(p)\big) = p.$$

The logit maps a probability to the real line; we call its value the *log-odds* or *evidence*. The decisive property for fusion is that **Bayesian updates are additive in log-odds**: multiplying odds corresponds to adding logits.

2.2 The exponential family and the canonical link

Definition 2.3 (exponential family). A distribution belongs to the exponential family if its density can be written $p(x \mid \theta) = h(x) \exp\big(\eta(\theta)\, T(x) - A(\theta)\big)$, with natural parameter η, sufficient statistic T, and log-partition A.

Proposition 2.4 (Bernoulli canonical link). *The Bernoulli distribution $p(y \mid \mu) = \mu^y (1 - \mu)^{1-y}$, $y \in \{0, 1\}$, is exponential-family with natural parameter $\eta = \operatorname{logit}(\mu)$ and mean function $\mu = \sigma(\eta)$.*

Proof. Write $p(y \mid \mu) = \exp\big(y \log \frac{\mu}{1-\mu} + \log(1 - \mu)\big)$. The coefficient of $T(y) = y$ is $\eta = \log \frac{\mu}{1-\mu} = \operatorname{logit}(\mu)$, whose inverse is $\mu = \sigma(\eta)$. □

This identifies log-odds as the *natural parameter space* of binary relevance and the sigmoid as the *canonical inverse link*. Both facts will recur structurally throughout the paper: every calibrated signal is a Bernoulli natural parameter, and every conversion back to probability is a sigmoid.

2.3 Bayesian inference for binary relevance

Let $R \in \{0, 1\}$ be the latent relevance of a document to a query, and let s be an observed score. By Bayes' theorem,

$$P(R = 1 \mid s) \;=\; \frac{P(s \mid R{=}1)\, P(R{=}1)}{P(s \mid R{=}1)\, P(R{=}1) + P(s \mid R{=}0)\, P(R{=}0)}.$$

Taking log-odds and writing $\pi = P(R{=}1)$ for the prior gives the additive form that organizes the whole paper:

$$\boxed{\operatorname{logit} P(R{=}1 \mid s) \;=\; \underbrace{\log \frac{P(s|R{=}1)}{P(s|R{=}0)}}_{\text{log-likelihood ratio (evidence)}} \;+\; \underbrace{\operatorname{logit}(\pi)}_{\text{prior}}} \tag{2.1}$$

Calibration is therefore reduced to specifying, for each signal, a model of its **log-likelihood ratio** $\Lambda(s) = \log \frac{P(s|R=1)}{P(s|R=0)}$. Section 3 chooses a parametric model for lexical scores; Section 4 estimates Λ nonparametrically for vector scores.

2.4 The Neyman–Pearson likelihood ratio

Definition 2.5 (likelihood ratio). For hypotheses $H_1 : R{=}1$ and $H_0 : R{=}0$ and observation s, the likelihood ratio is $\Lambda(s) = f_1(s)/f_0(s)$, where f_i is the density of s under H_i.

Theorem 2.6 (Neyman–Pearson). *The likelihood ratio is the canonical statistic for binary hypothesis testing: threshold rules on it are optimal in the Neyman–Pearson sense under fixed size, and the posterior depends on the observation through the likelihood ratio and the prior.*

Equation (2.1) is exactly Theorem 2.6 written in log-odds: the posterior is the log-likelihood ratio shifted by the prior log-odds and squashed by σ. The two calibration sections differ only in *how* Λ is modeled — parametrically from the score for BM25, empirically from a density ratio for vectors — and the fusion section builds on the additivity of (2.1): independent log-likelihood ratios sum, though, because real retrieval signals are correlated, the fusion of Section 5 applies a conservative pooling of this sum rather than the raw sum itself (Section 5.3).

---

3. Bayesian calibration of lexical scores

3.1 A sigmoid likelihood model

We model the log-likelihood ratio of a BM25 score as linear in the score. This is the minimal assumption consistent with "higher score ⇒ more evidence of relevance," and it makes the posterior a sigmoid.

Definition 3.1 (sigmoid likelihood). For a BM25 score s, model

$$\log \frac{P(s \mid R{=}1)}{P(s \mid R{=}0)} \;=\; \alpha\,(s - \beta),$$

where $\alpha > 0$ controls sensitivity (the slope of evidence in score) and β is the decision midpoint. Equivalently, the *likelihood term* used downstream is

$$L(s) \;=\; \sigma\Big(\alpha\,(s - \beta)\Big) \in (0, 1). \tag{3.1}$$

Under a neutral prior $\pi = \frac{1}{2}$, $L(s)$ is itself the posterior; with a non-trivial prior it enters (2.1) as the evidence term.

3.2 The base rate and the core posterior

Anchoring the sigmoid midpoint at the corpus median — the natural unsupervised choice for β — assigns $\approx 50\%$ relevance probability to scores near the median, even though most documents are irrelevant to a typical query. The result is systematic overconfidence, and no per-document feature can correct it, because the defect is a property of the corpus-level *prevalence* of relevance. A single corpus-level prior fixes it.

Definition 3.2 (base rate). The base rate $\pi_b \in (0, 1)$ is the corpus-level prior probability that a randomly chosen document is relevant to a randomly chosen query, independent of any document feature.

Theorem 3.3 (two-term posterior). *With likelihood term $L(s)$ (Def. 3.1) and base rate π_b, the calibrated posterior is a sum of two log-odds terms,*

$$\operatorname{logit} P(R{=}1 \mid s) \;=\; \operatorname{logit} L(s) + \operatorname{logit} \pi_b \;=\; \alpha\,(s - \beta) + \operatorname{logit} \pi_b. \tag{3.2}$$

Proof. Specialize the log-odds Bayes form (2.1) with the sigmoid-model log-likelihood ratio $\operatorname{logit} L(s) = \alpha(s - \beta)$ (Def. 3.1) and prior π_b. □

Equation (3.2) is the core of the lexical calibration: it carries no free constants beyond the two interpretable likelihood parameters α, β and the single scalar π_b.

Corollary 3.4 (neutrality). When $\pi_b = \frac{1}{2}$, $\operatorname{logit} \pi_b = 0$ and the posterior reduces to the bare likelihood $L(s)$.

Remark 3.5 (efficient computation). Equivalently, the posterior is one Bayes update in probability space, $P = L\pi_b \,/\, [\, L\pi_b + (1-L)(1-\pi_b)\,]$, avoiding transcendental calls on the hot path.

The base rate may be supplied or estimated **unsupervised** from the score distribution — a high percentile of per-query scores, the smaller component of a two-component score mixture, or the knee of the sorted-score curve. Crucially, π_b enters as a constant additive shift in log-odds, so it is a **monotone reparameterization** that does not change document order (Corollary 3.9); it changes only the absolute calibration of the output probabilities. Section 6.3 shows this single scalar removes 68–77% of the calibration error with no relevance labels.

3.3 An optional document-quality prior

The base rate is global. When static, per-document quality signals are available — length, term frequency, recency, source authority — they can be incorporated as an *independent* prior, appended as one further additive term.

Definition 3.6 (document-quality prior). Let $P_0(d) \in (0,1)$ be any prior probability of relevance computed from static document features. The augmented posterior is

$$\operatorname{logit} P(R{=}1 \mid s, d) \;=\; \alpha\,(s-\beta) + \operatorname{logit} \pi_b + \operatorname{logit} P_0(d).$$

Remark 3.7 (empirical Bayes, with a caveat). When P_0 is built from features that also enter the BM25 score, it is an *empirical-Bayes* prior (Robbins, 1956; Efron, 2010) rather than a classically independent one — informed by data through a different functional form than the likelihood, in the spirit of the James–Stein estimator. The specific functional form of P_0 is an application-dependent modeling choice and is **not** essential to the framework; we deliberately leave it unspecified, because hand-tuned feature-combination weights carry no principled justification, and empirically the unsupervised base rate of §3.2 already captures the calibration gain on its own — indeed the prior-free training mode of §3.6, which discards P_0 at inference, calibrates best in our experiments (§6.3). The document prior is therefore an optional extension; the core method is the two-term posterior (3.2). Unlike the base rate, a *document-dependent* prior can slightly perturb the ranking relative to BM25 by injecting a quality signal (Corollary 3.9).

3.4 Monotonicity and rank preservation

Theorem 3.8 (score monotonicity). *For any fixed prior, the calibrated posterior is strictly increasing in the BM25 score:* $s_1 > s_2 \Rightarrow P(R{=}1 \mid s_1) > P(R{=}1 \mid s_2)$.

Proof. For fixed prior p, $P(R{=}1 \mid s) = \sigma\big(\alpha(s-\beta) + \operatorname{logit} p\big)$ composes the strictly increasing affine map $s \mapsto \alpha(s-\beta)$ (as $\alpha > 0$) with the strictly increasing σ. □

Corollary 3.9 (rank invariance of the core posterior). *The two-term posterior (3.2) is a strictly monotone transform of the BM25 score; hence the base-rate calibration preserves the BM25 ranking exactly. An optional document-quality prior* $P_0(d)$ *(Def. 3.6) makes the posterior depend on both* s *and* d *and may therefore re-rank documents relative to BM25 — by design, since* P_0 *injects a quality signal.*

3.5 Compatibility with WAND and Block-Max WAND

Dynamic pruning skips documents that cannot enter the top-k. WAND (Broder et al., 2003) maintains, per query term, an upper bound on that term's score contribution; a document is skipped when the sum of its matched terms' upper bounds falls below the current top-k threshold. Block-Max WAND (BMW) (Ding and Suel, 2011) refines this with per-block maxima. Both are *exact*: they never skip a document that would have made the top-k.

Calibration must preserve this safety. A per-term BM25 upper bound exists because each saturating BM25 term is bounded by $\operatorname{IDF}(t)(k_1+1)$, giving a score upper bound U for any candidate.

Theorem 3.10 (exact pruning bound). *Under the core posterior (3.2), the probability is a strictly monotone transform of the score, so a BM25 score upper bound U maps directly to a probability upper bound*

$$P(R{=}1 \mid s) \;\le\; \overline{P}(U) \;:=\; \sigma\big(\alpha(U-\beta)+\operatorname{logit}\pi_b\big) \qquad \text{whenever } s \le U,$$

which is an exact WAND threshold in probability space; per-block maxima yield exact BMW bounds. If an optional document prior is used, replacing $P_0(d)$ by its maximum $\overline{P_0} = \max_d P_0(d)$ in the bound restores exactness, since the posterior is also monotone in the prior.

Proof. Monotonicity in the score (Theorem 3.8) gives $P(R{=}1 \mid s) \le \sigma(\alpha(U-\beta)+\operatorname{logit}\pi_b)$ for $s \le U$. With a document prior, monotonicity in the prior supplies the further bound at $\overline{P_0}$. □

Because calibration is a monotone transform, the ranking — and hence the set of documents the threshold admits — is identical to BM25's (Corollary 3.9), so pruning remains exact and calibration inherits the sub-linear query-time complexity of BM25 retrieval.

3.6 Parameter learning

The calibration parameters (α, β) — and, when desired, per-signal fusion weights — can be fit from relevance labels by minimizing binary cross-entropy.

Definition 3.11 (cross-entropy). For predictions $\hat{p}_i$ and labels $y_i \in \{0,1\}$, $\mathcal{L} = -\frac{1}{N}\sum_i \big[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)\big]$.

Theorem 3.12 (gradients). *Under the likelihood model $\hat{p}_i = \sigma(\alpha(s_i-\beta))$,*

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \frac{1}{N}\sum_i (\hat{p}_i - y_i)(s_i - \beta), \qquad \frac{\partial \mathcal{L}}{\partial \beta} = -\frac{\alpha}{N}\sum_i (\hat{p}_i - y_i).$$

These follow from the sigmoid derivative and the chain rule. Training admits three modes that differ in how the prior is treated: **(C1) balanced**, fit on the likelihood $\sigma(\alpha(s-\beta))$; **(C2) prior-aware**, fit on the full posterior including an optional document prior (§3.3); and **(C3) prior-free**, fit on the likelihood with inference prior fixed at $\frac{1}{2}$ — the best-calibrated mode in our experiments (§6.3). Online refinement uses SGD with EMA-smoothed gradients and Polyak averaging of (α, β) for stable inference under streaming feedback, optionally with an exponential-decay weighting that tracks concept drift in non-stationary relevance.

The same cross-entropy machinery calibrates *neural* re-ranker scores for inclusion in the fusion below, either parametrically by Platt scaling $\sigma(a\,z+b)$ or nonparametrically by isotonic regression (PAVA), so that any learned scorer also contributes a calibrated log-odds term.

---

4. Bayesian calibration of vector scores

4.1 The likelihood-ratio posterior

The lexical calibration of Section 3 fixes a *parametric* form for the log-likelihood ratio. Vector scores admit no such natural parametric form, but they do not need one: the likelihood ratio can be estimated directly from the distribution of distances. Let δ denote a query–document distance (cosine distance, $1-\cos$, or Euclidean), let $f_R(\delta)$ be the density of distances among *relevant* documents, and let $f_G(\delta)$ be the *global* (background) density of distances across the corpus.

Theorem 4.1 (vector calibration posterior). *Under the likelihood-ratio model with background density f_G and relevant density f_R, and corpus prior P_{base}, the posterior probability of relevance given distance δ is*

$$\boxed{P(R{=}1 \mid \delta) \;=\; \sigma\Big(\underbrace{\log\frac{f_R(\delta)}{f_G(\delta)}}_{\text{vector evidence}} \;+\; \operatorname{logit}\big(P_{\text{base}}\big)\Big).} \tag{4.1}$$

Proof. Specialize (2.1) with $P(\delta \mid R{=}1) = f_R(\delta)$ and $P(\delta \mid R{=}0) \approx f_G(\delta)$ (the background dominated by irrelevant documents). The log-likelihood ratio is $\log f_R(\delta)/f_G(\delta)$; adding the prior log-odds and applying σ gives (4.1). □

Equation (4.1) is **structurally identical** to the lexical posterior (3.2): a log-likelihood ratio plus a prior, squashed by σ. The only difference is that the lexical ratio is the parametric $\alpha(s-\beta)$ while the vector ratio is the empirical $\log f_R/f_G$.

4.2 Vector evidence as a log density ratio

Definition 4.2 (vector evidence). $e_{\text{vec}}(\delta) = \log \dfrac{f_R(\delta)}{f_G(\delta)}$.

The evidence is positive exactly when a distance is more typical of relevant documents than of the corpus at large, and negative when it is more typical of the background. It quantifies *how much more likely* this proximity is under relevance than by chance.

Remark 4.3 (pointwise mutual information). The vector evidence is the pointwise mutual information between the observed distance and the relevance event: $e_{\text{vec}}(\delta) = \log \frac{P(\delta, R=1)}{P(\delta)\,P(R=1)}$ up to the prior. Calibration thus measures the information a distance carries about relevance, not the distance itself — which is why no fixed geometric rescaling (Section 1.1) can succeed: the informativeness of a given distance depends on the corpus.

Remark 4.4 (distribution-free). Unlike the parametric sigmoid likelihood, (4.1) assumes no functional form for f_R or f_G; both are estimated empirically. The framework adapts automatically to the geometry of each embedding model and corpus.

4.3 The circularity problem and cross-modal conditional independence

Estimating f_R requires knowing which documents are relevant — the very quantity calibration is meant to infer. Estimating f_R from the top-k retrieved set without correction is circular: the nearest neighbors are *defined* by small distance, so their distance density is biased toward small distances regardless of true relevance.

The resolution is to import an *external* relevance signal that is conditionally independent of vector distance. We stress at the outset that the resulting estimator is **pragmatic**: the independence it assumes is typically only approximate (Remark 4.6), so the construction below is best read as a principled heuristic that the experiments (Section 6.4) validate empirically, not as a guarantee.

Assumption 4.5 (cross-modal conditional independence). Given true relevance R, the vector distance δ and an external signal z (e.g., a lexical BM25 probability, an alternative embedding model, or an index-derived density prior) are conditionally independent: $P(\delta, z \mid R) = P(\delta \mid R)\,P(z \mid R)$.

Under Assumption 4.5, the external signal's relevance estimate provides *importance weights* on the retrieved distances that break the self-referential loop: distances are reweighted by how relevant each document is *according to a different modality*, so the reweighted distance density estimates f_R without using distance to define relevance.

Remark 4.6 (robustness to violation). Assumption 4.5 is rarely exact — text-trained embeddings correlate with lexical overlap. But the weighted estimator degrades gracefully: residual correlation biases the density estimate toward the external signal's view of relevance, which is a *useful* bias when that signal is informative and a controlled one when it is not. Section 6.4 measures this empirically and finds the cross-modal estimator outperforms structurally independent priors despite violating the assumption, except when the external signal is adversarial.

4.4 Density estimation

Given retrieved distances $\delta_1, \dots, \delta_n$ with external importance weights $w_1, \dots, w_n$ (from a signal satisfying Assumption 4.5), we estimate f_R two ways.

Definition 4.7 (importance-weighted KDE). The nonparametric estimate is

$$\hat{f}_R(\delta) \;=\; \frac{\sum_{i=1}^{n} w_i\, K_h(\delta - \delta_i)}{\sum_{i=1}^{n} w_i}, \qquad K_h(u) = \frac{1}{h\sqrt{2\pi}} \exp\Big(-\frac{u^2}{2h^2}\Big),$$

a weighted Gaussian-kernel density. The bandwidth h follows a weighted Silverman rule, $h = \left(\frac{4}{3n}\right)^{1/5} \hat{\sigma}_w$, with $\hat{\sigma}_w$ the weighted standard deviation.

Proposition 4.8 (circularity breaking, idealized). *If Assumption 4.5 held exactly, $\hat{f}_R$ in Definition 4.7 would be a consistent estimator of f_R as $n \to \infty$ and $h \to 0$ with $nh \to \infty$, despite the retrieved set being distance-selected, because the weights w_i would carry relevance information from a modality independent of δ. In practice the assumption is approximate, so this states the estimator's idealized justification rather than a guarantee; Section 6.4 reports its empirical behavior, including a case where the external signal is adversarial and the estimator correctly degrades.*

The weighted KDE performs a *soft partition* of the retrieved set: documents the external signal deems relevant contribute their distances to $\hat{f}_R$ in proportion to their weight, so $\hat{f}_R$ concentrates where relevant distances actually lie.

Definition 4.9 (Gaussian-mixture alternative). Parametrically, model the top-k distance distribution as a two-component mixture — a relevant component $\mathcal{N}(\mu_R, \sigma_R^2)$ and a background component fixed to the corpus statistics — fit by EM with the E-step responsibilities *initialized* from the external relevance weights. The relevant component then supplies f_R. The fixed background component prevents the mixture from collapsing both components onto the dominant (irrelevant) mass.

The nonparametric KDE makes no shape assumption and suits multimodal relevant densities; the parametric GMM is smoother and cheaper to evaluate and suits approximately Gaussian relevant masses. The background density f_G is estimated *once* per corpus from index-construction distances and reused for all queries.

4.5 Estimation without external signals

When no second modality is available (pure vector search), three fallbacks recover most of the calibration quality, in decreasing order of strength:

1. **Distance-gap detection.** The sorted top-k distances often exhibit a gap separating a tight cluster of near neighbors from the bulk; documents inside the gap receive high weight. This is a self-contained surrogate for the external signal.
2. **Index density prior.** ANN index structure supplies a density-based prior (Section 4.7) usable as the importance weight.
3. **Multi-model cross-calibration.** Two or more embedding models calibrate each other, each serving as the other's conditionally independent signal under Assumption 4.5.

Remark 4.10 (degradation hierarchy). Estimation quality degrades gracefully: cross-modal weighting $>$ index density prior $>$ distance-gap detection $>$ naïve unweighted KDE. Even the weakest strategy improves on the geometric rescaling of Section 1.1, because it conditions on the empirical distance distribution rather than ignoring it.

4.6 Concentration of measure

Theorem 4.11 (background concentration). *In a D-dimensional embedding space with D large, the background distance density f_G concentrates sharply: the distances of random corpus documents to a fixed query cluster within $O(1/\sqrt{D})$ of their mean.*

Proof sketch. For embeddings on the unit sphere, the inner product of a fixed query with a uniformly random point has variance $O(1/D)$ (Vershynin, 2018, Ch. 3); cosine distance inherits the same concentration. □

Corollary 4.12 (background stability). *The sharp concentration of f_G makes the corpus-level background statistics (mean and variance of distance) a stable, low-variance calibration signal that need be computed only once and is robust to which documents are sampled.*

Concentration also explains an empirical regularity (Section 6.4): narrower KDE bandwidths slightly improve ranking, because f_R is concentrated in high dimensions and over-smoothing blurs the relevant mass into the background.

4.7 Index-aware statistics

The calibration draws its statistics from structures the ANN index *already maintains*, at negligible additional cost.

IVF (inverted file). An IVF index partitions the corpus into C cells by k-means. Construction yields, per cell, the population N_c, centroid, and intra-cell residual distance statistics; search yields query-to-centroid distances and per-cell candidate distances. These furnish both f_G (from corpus-wide residuals) and a cell-density prior. Sparse cells are more discriminative — proximity within a sparse region is stronger evidence — yielding an IDF-like prior

$$\pi_{\text{cell}} \;=\; \sigma\Big(\gamma\Big(\frac{\overline{N}}{N_c} - 1\Big)\Big), \tag{4.2}$$

with $\overline{N}$ the average cell population and γ a sensitivity scale; $\pi_{\text{cell}} > \frac{1}{2}$ when $N_c < \overline{N}$.

HNSW (hierarchical navigable small world). An HNSW graph (Malkov and Yashunin, 2018) exposes edge distances, per-node degree, and the search trajectory (the sequence of visited-node distances). Edge-distance distributions estimate local density; the trajectory's terminal distances estimate f_R near the query. A k-th-neighbor distance prior, $\pi_{\text{knn}} = \sigma\Big(\gamma(\overline{\delta}_{\text{med}}/\delta_{(k)} - 1)\Big)$, plays the role of (4.2) for graph indexes.

Remark 4.13 (the index as a statistical model). Both index families are, when read this way, nonparametric density estimators of the corpus geometry built for free during indexing. Calibration reuses that estimate rather than recomputing it, so the marginal cost of producing calibrated probabilities over raw scores is a handful of arithmetic operations per candidate.

---

5. Probabilistic fusion: the log-odds conjunction

5.1 The conjunction shrinkage problem

Once every signal yields a calibrated probability, the signals must be combined. The textbook rule for independent events is the product rule, but it is the wrong rule for *evidence aggregation*.

Definition 5.1 (product conjunction). For calibrated probabilities $p_1, \dots, p_n$, the product rule gives $P_\wedge = \prod_{i=1}^{n} p_i$ (assuming independence).

Theorem 5.2 (shrinkage). *If n signals all report the same probability $p \in (0, 1)$, then $P_\wedge = p^n \to 0$ as $n \to \infty$, and $P_\wedge$ is strictly decreasing in n.*

Adding a *second confirming* signal of probability 0.8 drops the conjunction from 0.8 to 0.64; a third drops it to 0.512. Agreement among informative signals should *increase* confidence, yet the product rule monotonically destroys it. The defect is semantic: the product rule answers "what is the probability that *all* of n independent events occur?" whereas fusion asks "given n pieces of evidence about *one* event, how probable is it?" These are different questions.

Proposition 5.3 (information loss). *The product rule retains only $\sum_i \log p_i$, discarding the distribution of the $\log p_i$ across signals; it cannot distinguish unanimous moderate evidence from one strong and one weak signal with the same product.*

5.2 Log-odds mean aggregation

The fix is to aggregate in the natural parameter space (log-odds), where Bayesian evidence is additive, and to *average* rather than *sum* so the result does not drift with the number of signals.

Definition 5.4 (log-odds mean). $\overline{\ell} = \dfrac{1}{n}\sum_{i=1}^{n} \operatorname{logit}(p_i)$.

Theorem 5.5 (scale neutrality). *If $p_i = p$ for all i, then $\overline{\ell} = \operatorname{logit}(p)$ and the aggregate probability $\sigma(\overline{\ell}) = p$, independent of n.* The log-odds mean neutralizes the signal-count dependence that causes shrinkage, preserving the average evidence.

Theorem 5.6 (equivalence to normalized Logarithmic Opinion Pooling). *The log-odds mean is exactly the normalized Logarithmic Opinion Pool (Log-OP) of the calibrated signals,*

$$\sigma(\bar{\ell}) \;=\; \frac{\big(\prod_i p_i\big)^{1/n}}{\big(\prod_i p_i\big)^{1/n} + \big(\prod_i (1-p_i)\big)^{1/n}},$$

i.e. the geometric mean of the odds, renormalized — equivalently Hinton's (2002) Product of Experts with uniform exponents $1/n$.

Proof. $\bar{\ell} = \frac{1}{n}\sum \log \frac{p_i}{1-p_i} = \log\big(\prod_i \frac{p_i}{1-p_i}\big)^{1/n}$; applying σ and clearing gives the renormalized geometric mean of odds. □

This is the key structural fact behind the fusion: **the aggregation is a Product of Experts in log-odds space**, linear in the logits, so combining calibrated experts reduces to summing their log-odds.

5.3 Confidence scaling and the $\sqrt{n}$ law

Averaging discards a legitimate effect: n independent signals that *agree* should warrant more confidence than one. We restore this by scaling the mean evidence by a power of n.

Definition 5.7 (log-odds conjunction). For an exponent $\rho \geq 0$,

$$P_{\text{fuse}} \;=\; \sigma\Big(n^{\rho}\,\bar{\ell}\Big) \;=\; \sigma\Big(n^{\rho-1}\sum_{i=1}^{n}\operatorname{logit}(p_i)\Big). \tag{5.1}$$

Theorem 5.8 (CLT-motivated scaling, $\rho = \frac{1}{2}$). *Setting $\rho = \frac{1}{2}$ gives $P_{\text{fuse}} = \sigma\big(\frac{1}{\sqrt{n}}\sum_i \operatorname{logit} p_i\big)$, a CLT-motivated compromise: aggregate evidence still grows with the number of agreeing signals, but at the $\sqrt{n}$ rate of a standard-deviation-scaled sum rather than the n rate that a full independence assumption would imply. It is the variance-stabilized middle between ignoring signal count ($\rho = 0$) and treating every signal as independent ($\rho = 1$).*

Theorem 5.9 (sign preservation). *The multiplicative scaling n^{ρ} preserves the sign of $\bar{\ell}$: if every signal reports irrelevance ($p_i < \frac{1}{2}$, $\operatorname{logit} p_i < 0$), then $\bar{\ell} < 0$ and $P_{\text{fuse}} < \frac{1}{2}$. No setting of ρ can invert unanimous evidence.* This is a structural guarantee of the multiplicative form — an additive conjunction bonus could flip the sign, but $n^{\rho}\bar{\ell}$ cannot.

Proposition 5.10 (single-signal identity). *When $n = 1$, $n^{\rho} = 1$ and $P_{\text{fuse}} = \sigma(\operatorname{logit} p_1) = p_1$: a single signal passes through unchanged.*

Why the average, and not the sum. A reader trained in Bayes might expect pure summation: for conditionally independent evidence, the posterior log-odds are the prior log-odds plus the sum of each signal's log-likelihood ratio. Summation is indeed the Bayes-optimal rule — but under two conditions that retrieval signals violate. First, it assumes the signals are conditionally independent given relevance; real signals are correlated (BM25 and a text-trained encoder both respond to lexical overlap), and summing correlated evidence double-counts the shared component, producing overconfidence. Second, it assumes the inputs are raw likelihood ratios, whereas our p_i are already calibrated posteriors that each absorb a prior, so summing their logits double-counts the prior as well. The normalized Log-OP — the *average* (Theorem 5.6) — is a conservative aggregation rule that does not require specifying a joint dependence model: Logarithmic Opinion Pooling is the distribution closest in weighted KL to the pooled experts, and the geometric mean of odds stays bounded when correlated experts agree, where a sum would diverge. It does not claim to recover the true posterior under dependence — it sidesteps the overcounting of summation without committing to a dependence structure. The exponent ρ then restores a *controlled* confidence gain for genuine agreement, interpolating between the regimes: $\rho = 0$ is pure pooling (maximally conservative, ignoring how many signals there are), $\rho = 1$ is pure summation (the fully-independent limit), and $\rho = \frac{1}{2}$ is the variance-correct setting for partially-dependent signals (Theorem 5.8). Pure summation is therefore not rejected but recovered as the $\rho = 1$ endpoint — appropriate when independence genuinely holds, and avoided as the default precisely because it usually does not.

Posterior logits versus evidence logits. One clarification reconciles the pooling above with the additive posterior of Section 5.6, since two different quantities can be aggregated. If a signal exposes a *posterior* p_i computed under its own prior π_i, its *evidence* contribution is the prior-removed logit $\ell_i = \operatorname{logit}(p_i) - \operatorname{logit}(\pi_i)$. Pure Bayesian fusion — valid when the signals are conditionally independent — sums evidence logits and adds the target prior exactly once,

$$P_{\text{fuse}} = \sigma\Big(\operatorname{logit}\pi + \textstyle\sum_i \big(\operatorname{logit}(p_i) - \operatorname{logit}(\pi_i)\big)\Big).$$

This is the form of the unified posterior of Section 5.6, in which each evidence term — $\alpha(s-\beta)$ for BM25, $\log f_R(\delta)/f_G(\delta)$ for the vector signal — is already a prior-free logit and the corpus prior enters once. Conservative Log-OP fusion instead pools the (optionally prior-subtracted) logits and applies the confidence scaling,

$$P_{\text{fuse}} = \sigma\Big(\operatorname{logit}\pi + n^{\rho}\textstyle\sum_i w_i\big(\operatorname{logit}(p_i) - \operatorname{logit}(\pi_i)\big)\Big),$$

which is the safe default when the conditional independence of Section 5.6 is only approximate, or when a signal exposes only a posterior score. Unless stated otherwise we fuse likelihood-only (evidence) logits, so per-signal priors are not double-counted; the experiments of Section 6 use the pooled form above.

Per-signal reliability weights $w_i \geq 0$ with $\sum_i w_i = 1$ generalize the uniform mean to a weighted Log-OP, $P_{\text{fuse}} = \sigma\big(n^{\rho}\sum_i w_i \operatorname{logit} p_i\big)$, recovering Definition 5.7 at $w_i = 1/n$. Weights may be fixed (Naïve-Bayes uniform), learned from labels by a backprop-free Hebbian gradient on the simplex, or made query-dependent (Section 5.4).

5.4 Query-adaptive fusion weights

The weights of Section 5.3 — uniform or globally learned — treat each signal's reliability as constant across queries. Yet the relative value of lexical and dense evidence is query-dependent: a query of rare, precise terms is well served by BM25, while a paraphrastic or conceptual query is better served by the dense signal. We therefore let the fusion weights depend on the query.

Definition 5.11 (query-adaptive fusion). Given query features $\phi(q) \in \mathbb{R}^m$ — for example query length, term rarity, or the score dispersion of each signal — the per-signal weights are produced by a softmax gate, and the evidence logits $\ell_i = \operatorname{logit}(p_i) - \operatorname{logit}(\pi_i)$ (Section 5.3) are fused as a query-weighted Log-OP:

$$w_i(q) = \operatorname{softmax}\big(W\,\phi(q) + b\big)_i, \qquad P_{\text{fuse}}(q) = \sigma\Big(\operatorname{logit}\pi + n^{\rho}\textstyle\sum_i w_i(q)\,\ell_i\Big),$$

with W and b learned from labels by gradient descent on cross-entropy. The weights adapt each signal's reliability per query while the aggregation remains a weighted Log-OP (Theorem 5.6); applying per-signal normalization to the ℓ_i before the gate keeps a heavy-tailed signal from dominating.

Remark 5.12 (multi-head and robustness). Running several gates with different initializations and averaging their log-odds before the final sigmoid — a multi-head variant — reduces the variance of the learned weighting and improves robustness, much as an ensemble does. This query-conditioned weighted Log-OP coincides with the attention mechanism of Vaswani et al. (2017), but the query-adaptive weighting stands on its own as a fusion technique and the results here do not depend on that connection.

Optional logit gating. A further variant applies a fixed nonlinearity to each evidence logit ℓ_i before aggregation — $\operatorname{ReLU}(\ell) = \max(0, \ell)$, $\operatorname{Swish}_\beta(\ell) = \ell\,\sigma(\beta\ell)$, the Gaussian-gated GELU, or softplus — to suppress weak or negative evidence. Such gating is useful only when most signals are noise to be filtered out; in a two-signal hybrid setting it discards informative evidence and underperforms, as Section 6.2 confirms.

5.5 Boolean operations

The same log-odds machinery supports Boolean query composition under independence: conjunction by (5.1), disjunction by the complement rule $P_\vee = 1 - \prod_i (1 - p_i)$ (computed in log-space for numerical stability), and negation by the complement $P_\neg = 1 - p$. Exclusion queries such as "python AND NOT java" compose as $P_\wedge(p_{\text{python}},\, 1 - p_{\text{java}})$. All operations are closed on calibrated probabilities and commute correctly with thresholding — unlike RRF (Section 1.2).

5.6 The unified hybrid posterior

Because every calibrated signal is a log-likelihood-ratio term and (2.1) is additive, fusing lexical and dense evidence is a single sum in log-odds.

Theorem 5.13 (unified posterior). *For a query q and document d with BM25 score s, vector distance δ, and corpus prior P_{base}, under cross-modal conditional independence (Assumption 4.5),*

$$\boxed{\operatorname{logit} P(R{=}1 \mid q, d) \;=\; \operatorname{logit}\left(P_{\text{base}}\right) \;+\; \underbrace{\alpha\,(s{-}\beta)}_{\text{lexical evidence}} \;+\; \underbrace{\log\frac{f_R(\delta)}{f_G(\delta)}}_{\text{vector evidence}}\;.} \tag{5.2}$$

Proof. Conditional independence factorizes the joint likelihood, $P(s, \delta \mid R) = P(s \mid R)\,P(\delta \mid R)$. Taking log-odds, the lexical term is the sigmoid-model log-likelihood ratio $\alpha(s - \beta)$ (Def. 3.1) and the vector term is the empirical log density ratio (Def. 4.2); the prior contributes $\operatorname{logit} P_{\text{base}}$. □

Every term here is an evidence (prior-free) logit and the prior appears exactly once; the posterior-versus-evidence distinction in Section 5.3 explains how this clean additive form relates to the pooled fusion used in practice when signals are correlated or expose only posterior scores.

Corollary 5.14 (extensibility). *The additive structure admits arbitrary further conditionally independent signals — graph centrality, temporal recency, behavioral signals — each appended as its own calibrated log-odds term e_k: $\operatorname{logit} P = \operatorname{logit} P_{\text{base}} + \sum_k e_k$. No existing signal's calibration needs to be re-fit when a new one is added; each is calibrated against the statistics of its own scoring mechanism or index. (A learned query-adaptive weighting, Section 5.4, would be re-trained to accommodate the new signal, but the per-signal calibrations are untouched.)*

Theorem 5.15 (structural unification). *BM25 and vector calibration instantiate one abstract pattern, $\operatorname{logit} P(R \mid s) = \log\big(\text{likelihood ratio of } s\big) + \operatorname{logit}(\text{prior})$, differing only in the distributional model of the observable: parametric ($\alpha(s - \beta)$) for lexical scores, empirical ($\log f_R / f_G$) for distances.* Sparse and dense retrieval are thereby the same Bayesian computation drawing on different native statistics — replacing the incommensurable scales of convex combination, the magnitude loss of RRF, and the query-dependent rescaling of min–max with a single principled fusion that requires no tuning beyond per-signal calibration.

---

6. Experiments

6.1 Setup

We evaluate on five BEIR datasets (Thakur et al., 2021): ArguAna, FiQA, NFCorpus, SciDocs, and SciFact, spanning argument retrieval, financial QA, biomedical citation, scientific document similarity, and scientific fact verification. The dense encoder is `all-MiniLM-L6-v2`; BM25 uses $k_1 = 1.2$, $b = 0.75$ with the Lucene variant and a Snowball English stemmer. Hybrid retrieval follows a retrieve-then-evaluate protocol: top-1000 candidates per signal, union, then re-scored. Calibration quality is measured by expected calibration error (ECE), Brier score, and log loss; ranking quality by NDCG@10, MAP@10, and Recall@10. Unless noted, all methods are **zero-shot**: they require no relevance labels.

6.2 Hybrid search ranking quality

Table 2 reports NDCG@10. The log-odds framework's best zero-shot configuration (balanced fusion with per-signal logit normalization and a query-dependent attention layer) is competitive with reciprocal rank fusion and convex combination on aggregate, while additionally producing calibrated probabilities, which RRF and convex combination do not.

Table 2. NDCG@10 on five BEIR datasets (zero-shot). Column maximum in bold.

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
RRF	39.61	36.85	34.43	20.11	71.43	40.49
Convex	**40.01**	37.10	35.60	19.67	73.37	41.15
Log-odds conjunction	0.92	32.59	35.31	18.44	72.00	31.85
Log-odds (local calib.)	39.79	37.19	34.10	19.51	**73.80**	40.88
Balanced fusion	37.27	**40.58**	**35.73**	21.42	72.47	41.50
Balanced + attention (norm.)	37.22	40.53	35.42	21.91	73.24	**41.67**
Multi-head attention (norm.)	37.13	39.05	35.70	21.78	70.59	40.85
Vector-calibrated + attention	37.66	39.81	34.82	**21.94**	71.34	41.11
Gated (ReLU)	35.16	27.54	32.45	17.08	69.01	36.25
Gated (Swish)	36.20	27.39	28.66	16.82	68.61	35.54
Probabilistic OR	0.06	25.52	33.46	15.89	66.95	28.38

Three observations follow.

Calibrated fusion beats rank fusion. The balanced-with-attention configuration (avg. 41.67) exceeds RRF (40.49) by $+1.18$ and edges convex combination (41.15), the strongest tuning-free baseline, while retaining probabilistic outputs. The $\sqrt{n}$-scaled log-odds conjunction with symmetric local calibration (40.88) already surpasses RRF without any per-signal weighting.

Failure modes are informative. The table deliberately includes methods that *underperform* BM25, because each failure validates a design choice:

- *Probabilistic OR* (avg. 28.38) collapses on ArguAna (0.06): the disjunction rule assumes independence and combines miscalibrated distributions without logit-space normalization, exactly the regime the log-odds conjunction is built to avoid (Section 5.1).
- *Gated fusion* (ReLU/Swish, avg. ≈ 36) underperforms in the two-signal hybrid setting because sparse gating (Section 5.4) is designed for high-dimensional signal spaces where most inputs are noise; with one sparse and one dense signal there is no noise to suppress, and the gate discards useful evidence.
- The *un-normalized log-odds conjunction* (31.85) collapses on ArguAna (0.92) because ArguAna's counter-argument retrieval makes BM25 an *adversarial* signal whose heavy-tailed logits dominate the fusion; per-signal logit normalization (the balanced variant) restores 37.27.

Vector calibration is competitive without labels. Replacing the geometric rescaling of the dense signal with the density-ratio calibration of Section 4 (vector-calibrated + attention, 41.11) is competitive with the best methods and obtains the strongest SciDocs result (21.94), confirming that index-derived distance statistics calibrate dense scores effectively with no supervision.

Table 3. Best zero-shot improvement over BM25 (NDCG@10).

Method	Δ vs BM25
Balanced + attention (norm.)	$+6.28$
Balanced fusion	$+6.11$
Convex	$+5.76$
Vector-calibrated + attention	$+5.73$
Log-odds (local calib.)	$+5.50$
RRF	$+5.11$
Dense	$+2.94$

MAP@10 and Recall@10 (omitted for space) track NDCG@10: balanced-with-attention attains the best aggregate MAP@10 (30.54 vs RRF 29.31, convex 30.07) and the best aggregate Recall@10 (49.93, essentially tied with convex 49.85 and RRF 49.89).

6.3 Probability calibration and the base rate

Ranking quality is unaffected by the base rate (Corollary 3.9), but calibration is transformed by it. Table 4 reports ECE before and after the base-rate correction; the corpus-level base rate reduces ECE by 68–77% with *no* relevance labels, and an explicit small base rate (or supervised batch fit) reaches near-perfect calibration.

Table 4. Expected calibration error (lower is better). Parenthetical = reduction vs. no base rate.

Method	NFCorpus ECE	SciFact ECE
Calibrated, no base rate	0.6519	0.7989
Calibrated, base rate = auto	0.1461 (-77.6%)	0.2577 (-67.7%)
Calibrated, base rate = 0.001	0.0081 (-98.8%)	0.0354 (-95.6%)
Supervised batch fit + base rate = auto	0.0085 (-98.7%)	0.0021 (-99.7%)
Prior-free training (C3)	0.0029 (-99.6%)	0.0058 (-99.3%)

The unsupervised auto-estimated base rate (a high score percentile, a mixture, or an elbow) captures most of the available calibration improvement; supplying the true corpus base rate or a handful of labels closes nearly all of the gap. Because the correction is a monotone log-odds shift, it never disturbs the ranking that produced these probabilities.

6.4 Vector calibration ablations

Conditional independence penalty. Section 4.3 calibrates the relevant-distance density using an external signal assumed conditionally independent of distance (Assumption 4.5). We compare a *structurally independent* density prior (IVF cell density / gap detection) against *cross-modal BM25 weights*, which violate the assumption because text embeddings correlate with lexical overlap. Table 5 reports the resulting hybrid NDCG@10.

Table 5. Conditional-independence penalty: density prior (CI-compliant) vs. BM25 weights (CI-violating).

Calibration weight source	ArguAna	FiQA	NFCorpus	SciDocs	SciFact	Average
Density prior (CI-compliant)	1.66	17.76	25.90	12.75	42.12	20.04
BM25 weights (CI-violating)	0.02	24.38	35.61	13.53	59.95	26.70

The CI-violating BM25-weighted estimator wins on four of five datasets (avg. 26.70 vs 20.04): the information gained from a cross-modal lexical signal dominates the bias from the dependence it introduces, validating the *pragmatic* use of Assumption 4.5 even when it does not hold exactly. The exception is decisive: on ArguAna the relationship reverses (1.66 vs 0.02), because counter-argument retrieval makes BM25 *adversarial*, and the calibration faithfully propagates the misleading weights into degraded probabilities. This bidirectional behavior is the desired property — the likelihood-ratio calibration amplifies informative cross-modal signal and exposes harmful signal rather than masking it.

Bandwidth. Scaling the Silverman KDE bandwidth by $c \in \{0.2, 0.5, 1.0, 2.0\}$ produces a small but consistent preference for *narrower* bandwidths ($c = 0.2$: avg. 29.16 over the four non-adversarial datasets, vs $c = 2.0$: 28.36), consistent with the concentration of f_R in high dimensions (Theorem 4.11): over-smoothing blurs the concentrated relevant mass into the background. The density ratio is otherwise robust to bandwidth choice.

Calibration baselines. Among monotone calibration transforms for the dense signal, the global-sigmoid (density-ratio) calibration is far better calibrated than arctangent normalization (ArguAna ECE 0.009 vs 0.186) and competitive with *supervised* Platt scaling — while requiring no labels.

6.5 Sparse-only calibration and threshold transfer

A calibrated probability supports a *fixed* relevance threshold that transfers across queries; a raw BM25 score does not, because its scale is query-dependent. We test this on NFCorpus and SciFact with queries split $50/50$: an F1-optimal threshold is selected on the training queries and applied *unchanged* to the held-out evaluation queries, so a small train$\rightarrow$test gap indicates a portable threshold. Because the base rate is a monotone reparameterization, the ranking — and hence the achievable F1 — is identical to BM25's (Corollary 3.9); what changes is whether a single threshold transfers.

Table 6. Threshold transfer: F1 of a train-selected threshold applied to held-out queries (train $\rightarrow$ test; gap $=$ train $-$ test, smaller is better). Recommended methods in bold.

Method	NFCorpus (train $\rightarrow$ test)	Gap	SciFact (train $\rightarrow$ test)	Gap
Calibrated, no base rate	0.1607 $\rightarrow$ 0.1511	0.0096	0.3374 $\rightarrow$ 0.2800	0.0574
Supervised fit, no base rate	0.1577 $\rightarrow$ 0.1405	0.0172	0.2358 $\rightarrow$ 0.2294	0.0064
Supervised fit + base rate (auto)	0.1559 $\rightarrow$ 0.1403	0.0156	0.3316 $\rightarrow$ 0.3341	-0.0025
Prior-free training (C3)	0.1808 $\rightarrow$ 0.1758	0.0050	0.2836 $\rightarrow$ 0.2852	-0.0016
Min–max rescaling	0.1796 $\rightarrow$ 0.1751	0.0045	0.3526 $\rightarrow$ 0.3486	0.0040
Platt scaling	0.0219 $\rightarrow$ 0.0193	0.0026	0.0005 $\rightarrow$ 0.0005	0.0000

Three readings. First, the **base rate sharply improves transfer**: without it the SciFact threshold loses 0.0574 F1 from train to test, whereas the base-rate-corrected threshold generalizes slightly *better* on held-out queries (gap -0.0025). Second, **a small gap alone is not sufficient** — Platt scaling has the smallest gaps, but its F1 collapses (to ≈ 0.02 on NFCorpus and ≈ 0 on SciFact), so its "transferable" threshold is meaningless; the base-rate and prior-free variants combine tight gaps with strong F1, and are also the best-calibrated methods of §6.3. Third, **min–max rescaling transfers thresholds too** with competitive F1, but it is a query-global rescaling rather than a calibration: it carries no probabilistic semantics and cannot supply the calibrated log-odds that the fusion of Section 5 consumes. Raw BM25 is absent because its unbounded, query-dependent scale admits no fixed threshold at all. Calibration — specifically the base rate — is what makes a single relevance threshold portable across queries, and it does so while preserving the BM25 ranking exactly.

6.6 Adoption

The calibrated operators have been taken up by independent systems. Apache Lucene merged a `BayesianScoreQuery` and `LogOddsFusionQuery` for probabilistic hybrid search; the Massive Text Embedding Benchmark (MTEB) added the calibrated scorer as a `bb25` retrieval baseline; the txtai engine added Bayesian BM25 as a hybrid score-normalization option (`normalize="bayesian-bm25"`); and Vespa.ai adopted it as an official sample application, with supporting platform changes exposing `averageFieldLength` as a rank feature. These integrations exercise the same log-odds posterior (Theorem 5.13) across separate codebases; the issues and pull requests are listed below so the claims can be checked directly.

Adoption references. Apache Lucene — #15827 (merged), #15948. MTEB — issue #4072, PR #4084. txtai — #1023 (PR #1037, milestone v9.6.0). Vespa.ai — vespa #36599; sample-apps #1930, #1869, #1873. Reference implementation — cognica-io/bayesian-bm25; an independent reproduction is instructkr/bb25.

6.7 Limitations

Our evaluation shows that calibrated log-odds fusion is competitive with strong tuning-free baselines while adding calibrated probabilities, but several caveats temper the ranking comparison and should be read alongside the numbers above.

1. **The ranking comparison is reported descriptively, without significance testing.** We report point estimates of NDCG@10, MAP@10, and Recall@10 without confidence intervals or query-level paired significance tests, so we describe the comparison against the strongest tuning-free baselines descriptively rather than asserting a *statistically established* improvement. We do not, however, treat the margins as negligible: the best configuration sits roughly a point of NDCG@10 above RRF on aggregate — a meaningful difference by retrieval standards — and is obtained while additionally producing the calibrated probabilities the baselines lack. More fundamentally, the contribution does not rest on this two-signal comparison: the substance of the framework is label-free calibration and the *additive fusion of an arbitrary number* of conditionally independent signals (Corollary 5.14), for which the sparse–dense benchmark is one validation rather than the headline result. The base-rate calibration gains (Section 6.3) rest only on a monotone reparameterization and are independent of all of these caveats.
2. **Single encoder and BM25 variant.** We use one dense encoder (`all-MiniLM-L6-v2`) and one BM25 configuration; stronger encoders or learned sparse retrievers may shift the absolute and relative numbers.
3. **No comparison with recent learned hybrid models.** We compare against classical fusion (convex, RRF) and calibration baselines, not against recent learned hybrid or late-interaction rerankers, which optimize ranking rather than calibration and would be the right comparison for a ranking claim.
4. **The vector calibration is a pragmatic estimator.** Its cross-modal conditional-independence assumption (Section 4.3) is typically violated; we report both its gains and its adversarial failure (ArguAna, Section 6.4) rather than claiming a guarantee.

A stronger evaluation — bootstrap confidence intervals, paired significance tests across queries, a documented hyperparameter-tuning protocol, and several encoder and BM25 variants — is the natural next step before asserting a ranking advantage over the tuning-free baselines.

---

7. Related work

Probabilistic retrieval. The probability ranking principle (Robertson, 1977) and the probabilistic relevance framework underlying BM25 (Robertson and Zaragoza, 2009) established relevance as a probabilistic event, but BM25 outputs an uncalibrated score; this work supplies the missing calibration and shows it preserves the ranking BM25 was designed to produce. Language-model and risk-minimization approaches (Lafferty and Zhai, 2001) model relevance probabilistically but do not address cross-signal calibration.

Score calibration and fusion. Platt scaling (Platt, 1999) and isotonic regression calibrate classifier scores from labels; our lexical and vector calibrations are unsupervised (base-rate and density-ratio estimation) and reduce to Platt scaling as the supervised special case (Section 3.6). Convex combination and RRF (Cormack et al., 2009) fuse signals without calibration and are subsumed, with their pathologies removed, by the log-odds posterior (Theorem 5.15).

Fusion and opinion pooling. Our log-odds fusion is a normalized Logarithmic Opinion Pool, equivalently a Product of Experts (Hinton, 2002); the query-adaptive weighting (Section 5.4) makes the pool's exponents depend on the query, which coincides with the attention mechanism (Vaswani et al., 2017). We use only the fusion algebra here and do not pursue that correspondence further.

Efficient inference. WAND (Broder et al., 2003) and Block-Max WAND (Ding and Suel, 2011) prune retrieval exactly; because our calibration is a monotone transform of the BM25 score, these algorithms apply unchanged and pruning remains exact (Section 3.5).

Vector indexing. IVF (Johnson et al., 2019) and HNSW (Malkov and Yashunin, 2018) are designed for fast approximate search; we read their construction-time statistics as a free nonparametric density model of the corpus geometry and reuse it for calibration. The concentration of measure that stabilizes the background density is standard in high-dimensional probability (Vershynin, 2018).

---

8. Conclusion

A single Bayesian principle calibrates every retrieval signal: the probability of relevance is the signal's log-likelihood ratio plus an independent prior, evaluated in the log-odds space that is the natural parameter of binary relevance. Instantiated for lexical scores, it gives a sigmoid-likelihood calibration whose base-rate term makes the posterior additive in log-odds — rank-preserving, prunable, and 68–77% better calibrated without labels. Instantiated for vector scores, it gives a likelihood-ratio calibration that reads its statistics from the ANN index for free and breaks the circularity of relevant-density estimation through cross-modal conditional independence. Because both are log-likelihood-ratio terms, fusion is a single additive posterior — a normalized Logarithmic Opinion Pool, equivalently a Product of Experts — that resolves the shrinkage pathology of naïve conjunction and unifies sparse and dense retrieval as one computation.

Empirically, on five BEIR datasets the framework is competitive with, and slightly exceeds on aggregate, the strongest tuning-free baselines — reciprocal rank fusion and convex combination (best zero-shot NDCG@10 $+6.28$ over BM25) — and unlike them it produces calibrated probabilities and admits exact dynamic pruning. We are deliberate about this comparison (Section 6.7): we report point estimates without significance testing and so do not assert a statistically established ranking win, though the aggregate margins are modest-to-meaningful rather than negligible. The durable contributions are label-free calibration and a fusion that scales to an arbitrary number of signals, not the two-signal ranking result. The same calibrated operators have been adopted across independent retrieval systems (Section 6.6).

Because the posterior is a sum of independent log-odds terms, it admits any number of further conditionally independent signals (Corollary 5.14): graph centrality, temporal recency, or behavioral feedback each calibrate against the statistics of their own scoring mechanism or index and append as one more log-odds term, with no re-calibration of the existing signals. The sparse–dense setting evaluated here exercises two such terms; the structure itself is indifferent to their number, which we regard as the central practical advantage over fusion schemes that must be re-tuned per signal.

References

1. Broder, A. Z., Carmel, D., Herscovici, M., Soffer, A., & Zien, J. (2003). Efficient query evaluation using a two-level retrieval process. *CIKM*.
2. Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. *SIGIR*.
3. Ding, S., & Suel, T. (2011). Faster top-k document retrieval using block-max indexes. *SIGIR*.
4. Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.
5. Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8), 1771–1800.
6. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.

7. Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. *SIGIR*.
8. Malkov, Y. A., & Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE TPAMI*, 42(4), 824–836.
9. Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*.
10. Robbins, H. (1956). An empirical Bayes approach to statistics. *Berkeley Symposium on Mathematical Statistics and Probability*.
11. Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, 33(4), 294–304.
12. Robertson, S. E., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389.
13. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *NeurIPS Datasets and Benchmarks*.
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *NeurIPS*.
15. Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.