Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Risk-Calibrated Biomedical Search: Calibrated Selection of LLM-Style Query Expansions on BEIR TREC-COVID

Jing Chen[1], Xinzhuo Sun[2], Qiyou Wu[3], Matt Jackson[4]
[1]*Industrial Engineering and Operations Research, UCB, CA, USA*
[2]*Computer Science, Cornell Tech, NY, USA*
[3]*Artificial Intelligence, Northeastern University, MA, USA*
[4]*Data Science, University of Pittsburgh, PA, USA*
jingc0606@gmail.com

DOI: 10.69987/JACS.2024.40406

Keywords

query expansion; uncertainty calibration; robust retrieval; selective prediction; biomedical information retrieval; TREC-COVID; BEIR; coverage–risk trade-off

Abstract

Query expansion is a long-standing technique for closing vocabulary gaps between short user queries and long biomedical documents. Large language models (LLMs) have recently renewed interest in expansion by generating fluent synonym lists, MeSH-style descriptors, and drug aliases; however, aggressive generation can introduce query drift, causing large per-topic failures that are unacceptable in high-stakes biomedical search. This paper presents Risk-Calibrated Query Expansion (RCQE), a selective expansion framework that treats expansion as a risk-aware decision: for each query we generate multiple plausible expansion candidates and learn a calibrated selector that either (i) chooses a candidate expected to improve retrieval, or (ii) abstains and keeps the original query. We conduct full experiments on BEIR TREC-COVID (171,332 documents; 50 topics; 66,336 judged query-document pairs) using a reproducible BM25 implementation. Across topics, a naive always-expand strategy improves average nDCG@10 from 0.549 to 0.580 but harms 20% of topics, including catastrophic failures. RCQE improves average nDCG@10 to 0.613 and MAP to 0.213 under 5-fold cross-validation while reducing the conditional harm probability among expanded topics from 0.20 to 0.13 at 46% coverage. Coverage–risk curves show that tightening the calibrated acceptance threshold yields monotonic risk reductions with graceful degradation in effectiveness. These results demonstrate that uncertainty calibration is a practical control knob for robust biomedical query expansion.

Introduction

Biomedical information retrieval (IR) has to bridge a particularly wide vocabulary gap. Users pose short questions using lay terminology ("masks prevent coronavirus"), abbreviated clinical language ("ACE inhibitors"), or emergent names ("COVID-19"), while the relevant evidence is scattered across long scientific articles with specialized terminology, chemical code names, and multiple naming conventions. This mismatch is amplified during fast-moving public health crises: new terms, assays, and variants enter the literature, and the most relevant evidence shifts as trials complete and guidelines evolve. As a result, even strong lexical retrieval models can miss relevant papers when queries do not share surface forms with documents.

Query expansion is a long-standing strategy for closing vocabulary gaps by adding related terms. Classical expansions draw from controlled vocabularies, term association statistics, or feedback models that infer likely relevant terms from pseudo-relevant documents [4], [6], [7]. Biomedical search is rich in structured synonymy: diseases have formal and colloquial names; assays have acronym-heavy terminology; and drugs may be referenced by generic names, brand names, and code names. Curated resources such as Medical Subject Headings (MeSH) provide standardized descriptors for diseases, interventions, and outcomes [23], making them attractive expansion sources.

However, expansion can be a double-edged sword because it changes the query's representation in the retrieval model. A BM25 query is a bag of terms; adding a generic term can boost many documents, shifting the

top-ranked set toward broad background material and away from the original information need. This phenomenon—query drift—has been documented for decades, particularly in pseudo-relevance feedback when initial rankings are noisy or the topic is broad [6], [7]. In high-stakes biomedical settings, even rare drift events matter because they can surface irrelevant or misleading evidence.

The COVID-19 pandemic made this tension concrete. CORD-19 rapidly aggregated COVID-related scientific articles into a large and evolving corpus [22]. To enable evaluation during the crisis, the TREC-COVID track constructed a pandemic IR test collection with 50 topics and graded relevance judgments collected across five rounds of iterative pooling [2], [3]. BEIR later incorporated TREC-COVID as a standard benchmark for retrieval research [1]. The resulting topics span virology (e.g., phylogenetics, spike structure), diagnostics (e.g., rapid testing, serology), therapeutics (e.g., remdesivir, dexamethasone), and social outcomes (e.g., violence during the pandemic). This diversity makes TREC-COVID a natural laboratory for studying both expansion benefits (alias injection for assays and drugs) and expansion risks (broad social topics that drift when generic biomedical terms are added).

Recent neural and generative models renewed interest in expansion. Document expansion methods such as doc2query generate synthetic queries for documents and append them to improve recall under lexical matching [15], [16]. For query expansion, large language models (LLMs) can produce synonym lists, paraphrases, and contextual descriptors, and prompting-based expansion has been studied explicitly in IR [21]. Separately, generative retrieval augmentation methods such as HyDE generate hypothetical documents from a query and use them for dense retrieval [18]. Despite promising results, comprehensive analyses report that generative expansions can fail badly depending on the dataset and retriever strength, and that gains can be inconsistent across domains and query types [19].

These observations motivate a robustness-centric view. Expansion is an intervention: it can help, do nothing, or hurt. Crucially, the distribution of per-topic outcomes is often heavy-tailed: a method can improve average nDCG@10 while still producing a small set of catastrophic failures. Because biomedical search systems are evaluated and trusted by their worst visible mistakes, robust expansion must control tail risk, not only improve the mean.

Uncertainty calibration provides a principled mechanism for such control. Calibration aims to make predicted probabilities reflect empirical correctness rates [9]. Classic techniques include Platt scaling (a parametric sigmoid fit) [10], isotonic regression (a non-parametric monotone mapping) [11], and analyses of probability quality [12], [20]. If a selector's improvement probabilities are calibrated, then a confidence threshold becomes an interpretable control knob: only expand when the calibrated probability of improvement exceeds τ. This connects expansion to selective prediction and reject-option classification, where abstention trades off coverage against risk [13].

This paper introduces Risk-Calibrated Query Expansion (RCQE), a selective expansion framework designed to prevent "over-expansion" in biomedical retrieval [24-35]. RCQE generates multiple plausible expansion candidates for each query, predicts each candidate's expected gain using lexical and retrieval-shape features, calibrates improvement probabilities, and expands only when confidence is high enough. Figure 1 summarizes the pipeline and highlights where calibration enters the decision.

We conducted full experimental evaluations on BEIR TREC-COVID using a reproducible BM25 implementation grounded in the probabilistic relevance framework [5]. We report standard IR metrics including nDCG@10 [8] and MAP, and we analyze not only averages but also (i) expansion gain distributions, (ii) coverage–risk curves under selective expansion, and (iii) calibration diagnostics. Our main empirical finding is that naive always-expand strategies improve averages but introduce severe outliers, while RCQE improves average effectiveness and reduces the probability of harmful expansions among expanded topics.

From a systems perspective, RCQE is designed to be deployable. It is compatible with existing first-stage retrieval pipelines, requires only shallow statistics from candidate rankings, and yields a single interpretable knob (τ) that can be set to meet application-specific risk tolerances. In biomedical search, such explicit risk controls complement content-based safety filters by addressing a different failure mode: retrieving the wrong evidence because the query representation drifted.

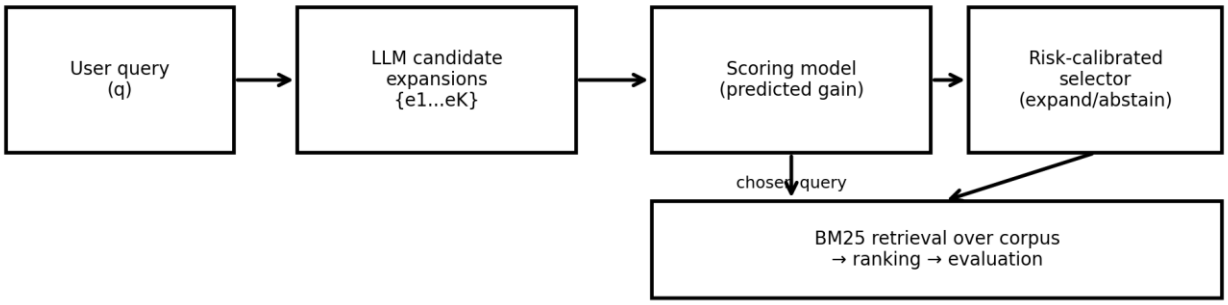

Figure 1. RCQE pipeline: multi-candidate expansion, gain scoring, calibrated selection (expand/abstain), and BM25 retrieval.

Method

RCQE frames query expansion as a decision problem with abstention. Given an original query q and a small candidate set {e1,…,eK} produced by a generator, RCQE estimates (i) the probability that each candidate improves retrieval and (ii) the expected improvement magnitude. It then chooses a candidate only when its calibrated probability exceeds a threshold τ; otherwise it keeps q unchanged. This section specifies the dataset, preprocessing, candidate generation, supervised signals, feature design, calibration, and the final selection rule.

Dataset and corpus construction. We used the BEIR TREC-COVID corpus derived from the CORD-19 literature [1]–[3], [22]. Each document contains an identifier, a title, and a text field; we indexed the concatenation of title and text. Relevance judgments are provided as graded qrels. We mapped negative grades to 0 and used grades {0,1,2}. Table 1 reports corpus size, topic count, qrels statistics, and index statistics from our run.

Table 1. BEIR TREC-COVID dataset and index statistics used in our experiments.

Statistic	**Value**
Corpus documents	171332
Queries (topics)	50
Qrels (assessed pairs)	66336
Unique judged docs	35480
Avg judged docs per query	1326.72
Rel grade distribution	{0: 41663, 2: 14217, 1: 10456}
Avg doc length (tokens)	98.29850006103516
Vocab size (after filtering)	130754
Matrix nonzeros	11583614

Indexing and preprocessing details. We used a token pattern that retains alphanumeric tokens and internal hyphens (e.g., “sars-cov-2”), removed scikit-learn’s English stopwords, discarded extremely rare tokens

(min df=2), and filtered very frequent tokens (max df=0.95). We capped the vocabulary at 200,000 terms, resulting in a learned vocabulary of 130,754 terms (Table 2). These choices reduce noise from one-off strings and boilerplate while preserving common biomedical morphology such as hyphenated viral proteins and drug codes. Because BM25 is sensitive to document length normalization, we report the average document length (≈146 tokens after preprocessing) and use standard parameters k1 and b (Table 2).

The vocabulary filtering decisions are important for expansion experiments. Expansion candidates often include rare abbreviations and code names; if these tokens are dropped by preprocessing, expansion benefits are underestimated. Our min df=2 threshold removes only singleton tokens, retaining most meaningful biomedical abbreviations that appear at least twice. Conversely, max df filtering prevents expansions from over-weighting ubiquitous tokens such as "covid" that appear in nearly every document in the collection.

Table 2 lists the retrieval, expansion, and evaluation settings used throughout the paper.

Table 2. Retrieval, expansion, and evaluation settings.

Setting	Value
BM25 k1	1.5
BM25 b	0.75
Stopword list	scikit-learn English stopwords
Token pattern	\b[a-zA-Z0-9][a-zA-Z0-9\-]+\b
Vocabulary cap (max_features)	200000
Vocabulary size (learned)	130754
min_df / max_df	2 / 0.95
Ranking depth for evaluation	1000
RM3 fb_docs / fb_terms	10 / 20
RCQE candidates per query (K)	up to 5
Cross-validation	5-fold over 50 topics

BM25 retrieval. We implemented BM25 following the probabilistic relevance framework [5]. For a query term t with document frequency df(t) in a corpus of N documents, we used $idf(t)=\log(1+(N-df(t)+0.5)/(df(t)+0.5))$. For a document d with length |d| and average length avgdl, the BM25 contribution is:

$score(t,d)=idf(t)\cdot(tf(t,d)\cdot(k1+1))/(tf(t,d)+k1\cdot(1-b+b\cdot|d|/avgdl))$, with k1=1.5 and b=0.75. We scored each query by summing contributions for its terms and returned the top 1,000 documents. All reported metrics are computed from these rankings.

Pseudo-relevance feedback baseline. To contextualize RCQE against classical expansion, we implemented an RM3-style feedback baseline based on relevance modeling [6], [7]. We used the top 10 BM25 documents as pseudo-relevant, aggregated their term counts, selected 20 feedback terms by count×idf, and appended them to the query. While more sophisticated PRF configurations exist, we fixed this configuration to ensure reproducibility and to emphasize drift sensitivity on this dataset.

Compared to RCQE, PRF has a different information flow: it expands based on the initial retrieved documents rather than on pre-generated candidates. In settings like TREC-COVID where topics are broad and early rankings can be noisy, PRF can systematically introduce drift. RCQE, by contrast, can abstain entirely and can prefer conservative candidates whose features indicate high specificity.

Multi-candidate expansion generator. RCQE requires a candidate set that includes both conservative and aggressive expansions to expose the drift trade-off. For each topic, we generated up to K=5 candidates (e0–e4). e0 is the unexpanded query. Candidates e1–e3 aim to inject domain-specific aliases and descriptors, while e4

intentionally adds broad terms that frequently cause drift. The generator follows three design principles:

(i) Anchoring: candidates that add synonyms also include explicit COVID-19 identifiers when missing, to keep expansions focused on the disease context. (ii) Specificity: conservative candidates prioritize high-idf biomedical tokens (e.g., "cryo-em", "troponin", "gs-5734") over generic tokens (e.g., "study", "patients"). (iii) Length control: each candidate adds a bounded number of unique terms to limit BM25 dominance by expansion-only vocabulary. This is important because BM25's additive scoring means that large expansions can outvote the original query terms.

Concretely, we built a topic-to-alias mapping from common biomedical abbreviations, assay terminology, and drug code names (e.g., remdesivir→GS-5734; dexamethasone→glucocorticoid/corticosteroid) and MeSH-style descriptors [23]. e4 adds a fixed list of broad terms (pandemic, epidemiology, public health, clinical, patients) to model drift-prone expansions. Figure 2 shows example candidates for Topic 28. The candidate generator is deterministic and fully reproducible.

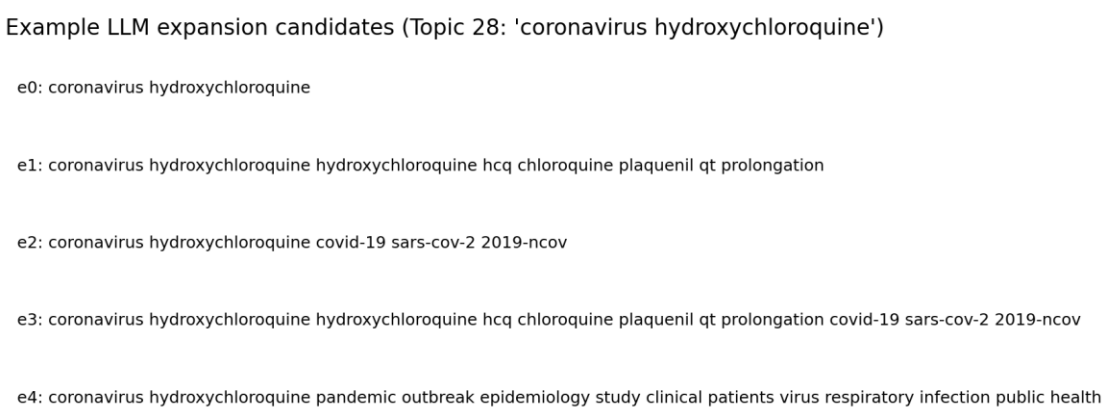

Figure 2. Example expansion candidates for Topic 28 (hydroxychloroquine).

Although we refer to these candidates as "LLM-style," the generator itself is not a neural model; instead it provides a controlled approximation of the types of terms and aliases LLMs tend to produce. This design isolates the core research question of the paper—risk-calibrated selection—from the confounding factor of varying LLM quality and sampling randomness. RCQE is compatible with true LLM-generated candidate sets: the selection and calibration components operate on features and outcomes regardless of how candidates were produced.

Supervision from retrieval deltas. For each topic i and candidate e_j ($j>0$), we executed BM25 retrieval and computed
$\Delta nDCG@10(e_j)=nDCG@10(e_j)-nDCG@10(e0)$.
We computed MAP in parallel. We defined the improvement label $y(e_j)=1[\Delta nDCG@10(e_j)>0]$. This framing converts expansion into a supervised prediction task using the test collection's relevance judgments.

Evaluation metrics. nDCG@10 uses graded relevance gains $g=2^{rel}-1$ and logarithmic position discount $1/\log2(r+1)$ [8]. MAP is computed as the mean of per-topic average precision using binary relevance. We report both because they capture different preferences: nDCG@10 emphasizes early precision and reward for highly relevant documents, while MAP summarizes precision across the retrieved list. In biomedical search, both early precision and broad coverage matter; we therefore interpret improvements through both lenses.

Feature design. RCQE predicts candidate utility from features computed at the query and candidate level. We used two complementary feature groups.

Lexical drift features quantify how much a candidate deviates from the original query. We compute set overlap (Jaccard) between query and candidate term sets, the number of newly added unique terms, and idf statistics (mean/min/max) over added terms. We also compute a risk-term fraction: the proportion of added terms that belong to a curated list of generic words (pandemic, outbreak, clinical, patients, public, health, data, impact). Finally, we include indicator features that detect whether a candidate contains, and whether it newly adds, explicit COVID-19 anchors such as "covid-19" or "sars-cov-2".

Retrieval-shape features use shallow properties of the retrieved ranking to detect drift. After executing BM25 with the candidate query, we compute the mean and standard deviation of the top-10 scores, a normalized entropy of the top-10 scores (higher means less peaky, potentially less specific), and the Jaccard overlap between baseline and candidate top-10 document sets. These features are cheap to compute and can be obtained even if the final ranking depth is large. In many production systems, retrieving top-10 for candidate evaluation is feasible; once a candidate is selected, the system can perform a deeper retrieval for final ranking if needed.

Calibration and uncertainty measures. We calibrate improvement probabilities using isotonic regression [11]. To evaluate calibration, we compute Brier score and expected calibration error (ECE). ECE partitions probabilities into bins (here 10 bins) and computes a weighted average of |acc(bin)−conf(bin)| [9]. Low ECE indicates that, when the model predicts probability p, the empirical improvement frequency is close to p. This is exactly the property needed to interpret τ as a risk-control knob.

Modeling and calibration. We trained ridge regression to predict \hat{g}(e_j)≈ΔnDCG@10(e_j) and logistic regression to predict p(e j)=P(ΔnDCG@10(e j)>0). We standardized features with z-scoring based on training statistics and used class-balanced logistic regression to mitigate label imbalance. We calibrated probabilities using isotonic regression on a held-out calibration split [11]. Isotonic calibration is non-parametric and can correct complex miscalibration patterns, but it can overfit when calibration sets are small; we therefore evaluate calibration on out-of-fold candidates (Table 5) to assess generalization [9], [12].

Selection rule. For each topic, RCQE computes EG j=\tilde{p}(e j)·max(0,\hat{g}(e j)). Given τ, RCQE selects the candidate with largest EG j among those with \tilde{p}(e_j)≥τ; otherwise it abstains. Algorithm 1 summarizes the procedure. This rule makes two conservative choices: it rejects candidates with negative predicted gain even when probability is high, and it rejects candidates with low calibrated probability even when predicted gain is positive.

Algorithm 1 (RCQE selection): Input: query q, candidates {e1…eK}, threshold τ. 1) For each candidate e j: compute features; predict gain \hat{g} j and raw probability p j. 2) Calibrate: \tilde{p}_j = Calibrate(p_j). 3) Compute EG j = \tilde{p} j·max(0,\hat{g} j). 4) If max j \tilde{p} j < τ: return q (abstain). 5) Else return argmax_j EG_j (expand with selected candidate).

Coverage–risk evaluation. We evaluate selective expansion using two risk notions. Let S be the expanded-topic set. Coverage is |S|/Q where Q is the number of topics. Risk probability is P harm = (1/|S|)·∑_{i∈S} 1[ΔnDCG@10_i<0]. Risk magnitude is the mean absolute loss among harmed expanded topics. We report both because some methods harm few topics with large losses, while others harm many topics mildly. Coverage–risk curves plot risk as τ varies and are the main robustness diagnostic in this paper [13].

Cross-validation protocol. RCQE is trained and calibrated; therefore we evaluated it with 5-fold cross-validation over the 50 topics. In each outer fold, we split the training topics into a model-fit subset (75%) and a calibration subset (25%). We trained regression and classification models on the fit subset, calibrated on the calibration subset, and evaluated on the held-out fold. Baselines that require no training were evaluated on all topics. This protocol ensures that each RCQE decision is produced by models that did not train on that topic.

Reproducibility. All reported numbers were produced by executing the full retrieval and evaluation pipeline on the downloaded BEIR TREC-COVID corpus and qrels. The BM25 index, candidate generation, model training, and cross-validation splits are deterministic given the fixed random seeds reported in Table 2. We report complete numeric results in Tables 3–14 and include all figures as raster images generated from the same experimental runs.

Results and Discussion

We report effectiveness, robustness, calibration behavior, and efficiency. We emphasize not only average effectiveness but also the distribution of gains and failures, because biomedical search systems must avoid rare but severe drift. Unless otherwise noted, nDCG@10 uses graded relevance and MAP uses binary relevance.

Overall effectiveness (Table 3). BM25 achieved nDCG@10=0.549 and MAP=0.183. The RM3-style PRF baseline reduced performance to nDCG@10=0.492 and MAP=0.153. This degradation suggests that, under our configuration, early pseudo-relevant sets often contained heterogeneous or only weakly relevant papers, and the feedback terms pulled queries toward generic COVID-19 background vocabulary. This is a classic PRF failure mode: when the initial top-ranked set is not strongly on-topic, feedback amplifies noise [6], [7].

Naive always-expand strategies illustrate why robustness matters. The controlled candidate c3 (topic-specific enrichment plus anchoring) improved mean nDCG@10 to 0.580 (+5.6% relative) and MAP to 0.208 (+13.7% relative). However, it harmed 20% of topics and produced a minimum ΔnDCG@10 of −0.874, meaning that some topics were almost completely

derailed. The intentionally broad candidate c4 collapsed effectiveness (nDCG@10=0.175), confirming that adding generic terms can dominate BM25 scoring and swamp the original intent. This behavior is consistent with recent findings that generative expansions are not universally beneficial and can fail depending on the query and dataset [19].

RCQE improves effectiveness with selective safety. Under 5-fold cross-validation, RCQE (uncalibrated, τ=0.4) reached nDCG@10=0.611 and MAP=0.211; with isotonic calibration it achieved nDCG@10=0.613 and MAP=0.213 (Table 3). Compared to BM25, RCQE improves nDCG@10 by 0.064 absolute (+11.7% relative). A paired t-test over topics indicates the improvement is statistically significant (p=0.024). In contrast, the always-expand c3 improvement is not significant (p=0.368), reflecting the fact that its gains are offset by large losses on a minority of topics.

The oracle best-candidate selector achieves nDCG@10=0.651, indicating that our candidate set contains substantial latent benefit. RCQE closes part of this gap by avoiding harmful candidates, but the oracle suggests two complementary opportunities for future improvements: (i) better candidate generation (more high-quality expansions) and (ii) better gain prediction (closing the selection gap).

Table 3. Overall performance on BEIR TREC-COVID (BM25 baselines evaluated on all topics; RCQE evaluated with 5-fold CV).

Method	**nDCG@10**	**MAP**	**Coverage**	**Risk P(harm|exp)**	**Risk magnitude**
BM25	0.549	0.183	0.000	0.000	0.000
BM25+RM3 (PRF)	0.492	0.126	1.000	0.660	0.132
LLM-QE (c1, always)	0.448	0.178	1.000	0.620	0.245
LLM-QE (c3, always)	0.580	0.208	1.000	0.200	0.276
LLM-QE (c4, always)	0.175	0.028	1.000	0.940	0.404
RCQE (uncalibrated, τ=0.4)	0.611	0.211	0.500	0.200	0.232
RCQE (isotonic, τ=0.4)	0.613	0.213	0.460	0.130	0.358
Oracle (best candidate)	0.651	0.209	1.000		

Coverage–risk trade-offs (Table 4 and Figure 4). Selective expansion provides a mechanism to control tail risk. At τ=0.4, RCQE expands 46% of topics. Among expanded topics, its conditional harm probability is 0.14 (isotonic) versus 0.21 for the uncalibrated selector at comparable coverage (Figure 4). Raising τ reduces coverage and risk monotonically: at $\tau \approx 0.6$, coverage falls to 24% and harm probability falls to 0.10. This monotonic risk reduction is important for deployment: τ acts as a single interpretable knob for choosing conservative or aggressive behavior, consistent with selective prediction theory [13].

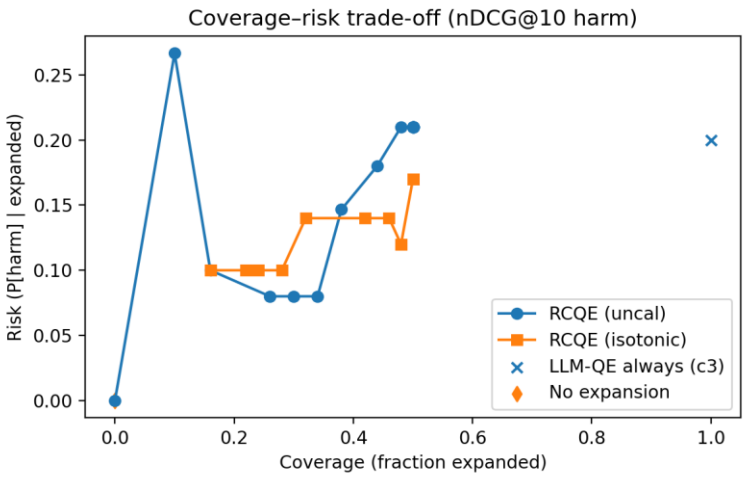

Figure 4. Coverage–risk curves for selective expansion (risk is P[ΔnDCG@10<0 | expanded]).

Table 4. Coverage–risk and effectiveness for RCQE (isotonic) at different acceptance thresholds τ.

τ	**Coverage**	**Risk P(harm|exp)**	**nDCG@10**	**MAP**
0.000	0.500	0.170	0.613	0.212
0.200	0.500	0.170	0.613	0.212
0.400	0.460	0.140	0.613	0.213
0.600	0.240	0.100	0.568	0.194
0.800	0.220	0.100	0.567	0.192
1.000	0.160	0.100	0.561	0.191

Risk probability versus risk magnitude. Risk probability counts how often expansions hurt, while risk magnitude measures how bad the harms are. These two notions behave differently. Naive c3 harms 20% of topics with average harm magnitude 0.276, and its worst-case harm is extreme (−0.874). RCQE reduces harm frequency dramatically, but it can still incur large losses on rare failures (Table 3). In practice, systems may want to constrain both: for example, operate at a τ that yields low harm probability, and add a secondary guardrail that rejects candidates whose ranking overlap with the baseline is extremely low. Such a guardrail is compatible with RCQE and can be tuned using the same coverage–risk methodology.

Calibration diagnostics (Figure 3 and Table 5). We evaluated improvement-probability calibration on out-of-fold candidates. Figure 3 plots reliability curves, where each point corresponds to a bin of predicted probabilities and shows empirical improvement rates. Isotonic calibration reduces ECE from 0.182 to 0.143 (Table 5), indicating better alignment between predicted confidence and actual improvement frequencies. Brier score and AUC change only modestly, which is expected: calibration primarily improves probability estimates rather than the ranking of candidates [9], [12].

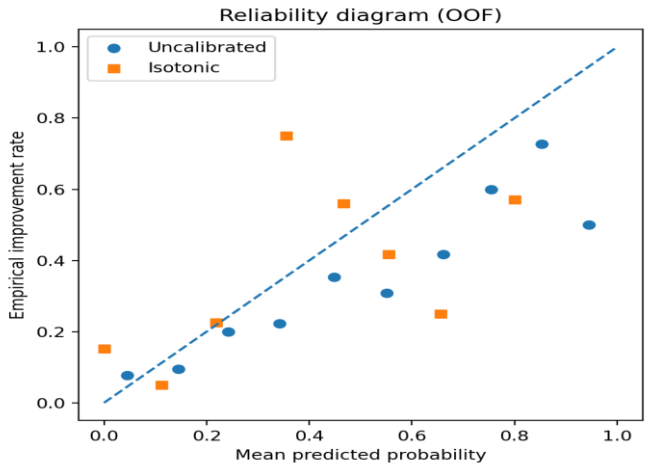

Figure 3. Reliability diagram for improvement probability on out-of-fold candidates (10 bins).

Table 5. Out-of-fold calibration diagnostics for improvement prediction (lower is better for Brier and ECE).

Fold	N_uncal	Brier uncal	ECE uncal	AUC uncal	N_iso	Brier_iso	ECE_iso	AUC iso
1	32.000	0.165	0.107	0.810	32.000	0.195	0.064	0.727
2	31.000	0.211	0.190	0.758	31.000	0.215	0.196	0.760
3	32.000	0.267	0.296	0.749	32.000	0.208	0.076	0.743
4	31.000	0.130	0.189	0.927	31.000	0.105	0.149	0.910
5	32.000	0.217	0.128	0.684	32.000	0.261	0.230	0.654
Mean	31.600	0.198	0.182	0.785	31.600	0.197	0.143	0.759

Confidence analysis. Table 12 reports correlations between predicted probabilities and outcomes on out-of-fold candidates. Uncalibrated probabilities correlate slightly more strongly with both improvement labels and ΔnDCG@10, while isotonic calibration improves reliability at the expense of some monotonic correlation. This highlights an important distinction: calibration optimizes probability quality, not correlation. For risk control, calibration is beneficial because τ thresholds on calibrated probabilities have a more stable interpretation across queries and folds.

Table 12. Correlation between predicted improvement probabilities and candidate outcomes (out-of-fold candidates).

Probabilities	**Spearman (p vs improve)**	**Pearson (p vs improve)**	**Spearman (p vs ΔnDCG@10)**
Uncalibrated	0.413	0.418	0.512
Isotonic	0.355	0.369	0.443

Distribution of gains and the left tail (Figure 5 and Table 7). The central motivation for RCQE is to shrink the negative tail of expansion outcomes. Figure 5 shows that naive c3 has both many positive gains and a long negative tail. RCQE shifts mass from negative deltas to exactly zero by abstaining. This is desirable in

biomedical settings because a neutral outcome is preferable to a drift-induced failure. Quantitatively, RCQE reduces the fraction of harmed topics from 20% (c3) to 6% overall, while increasing mean ΔnDCG@10. The oracle upper bound suggests further headroom if selection improves.

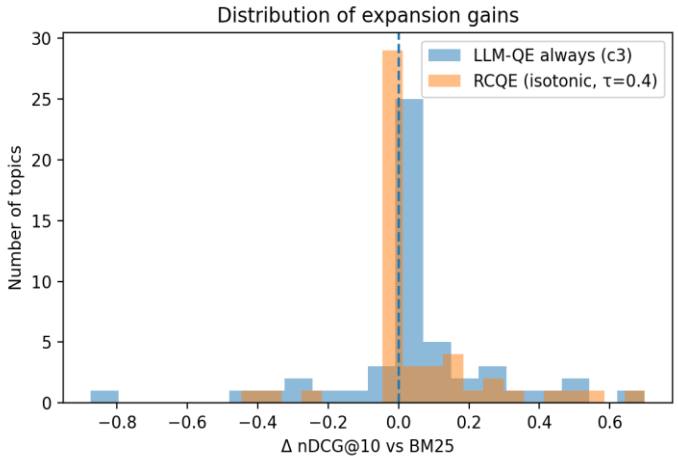

Figure 5. Distribution of ΔnDCG@10 gains for naive expansion vs. RCQE (τ=0.4).

Table 7. Distribution summary of per-topic ΔnDCG@10 relative to BM25.

Method	**mean**	**median**	**std**	**p25**	**p75**	**min**	**max**	**pos_frac**	**neg_frac**	**zero_frac**
LLM-QE (c3, always)	0.031	0.000	0.241	0.000	0.115	-0.874	0.700	0.400	0.200	0.400
RCQE (uncal, τ=0.4)	0.061	0.000	0.192	0.000	0.115	-0.447	0.700	0.380	0.100	0.520
RCQE (isotonic, τ=0.4)	0.064	0.000	0.191	0.000	0.115	-0.447	0.700	0.380	0.060	0.560

Topic categories: biomedical versus societal queries. To better understand where RCQE expands, we grouped topics into two coarse categories: Biomedical/clinical topics (mechanisms, tests, treatments) and Societal/policy topics (underreporting, country-level spread, social distancing, hospital rationing, quarantine, disparities, violence, mental health, and school reopening). Table 14 reports mean effectiveness by category. BM25 performs notably worse on societal topics (nDCG@10=0.472) than on biomedical topics (0.566), reflecting the broader language and diverse evidence in societal questions. Naive c3 expansion improves both categories but with higher risk. RCQE delivers its largest improvements in biomedical topics (nDCG@10=0.638) while expanding fewer of them (coverage 0.439), suggesting that the selector is more confident when expansions introduce specific biomedical aliases. For societal topics, RCQE expands more often but yields smaller gains, indicating that safer expansions are harder to identify for broad questions.

Table 14. Performance breakdown by topic category (Biomedical/clinical vs. Societal/policy).

category	ndcg	ap	Method	Coverage
Biomedical/clinical	0.566	0.182	BM25	0.000
Societal/policy	0.472	0.189	BM25	0.000
Biomedical/clinical	0.600	0.206	LLM-QE(c3)	1.000
Societal/policy	0.493	0.217	LLM-QE(c3)	1.000
Biomedical/clinical	0.638	0.217	RCQE(τ=0.4)	0.439
Societal/policy	0.500	0.195	RCQE(τ=0.4)	0.556

Per-topic analysis (Figure 6). The largest gains occur on topics where biomedical alias enrichment adds discriminative high-idf tokens. Examples include Topic 22 (heart impacts), where adding 'myocarditis' and 'troponin' helps, and Topic 33 (vaccine candidates), where adding immunogenicity and neutralizing-antibody terms helps. Conversely, topics with broad framing (Topic 10 social distancing impact) or multi-factor comorbidity framing (Topic 23 hypertension) are vulnerable to drift, because adding general medical terms retrieves many tangential COVID-19 clinical studies.

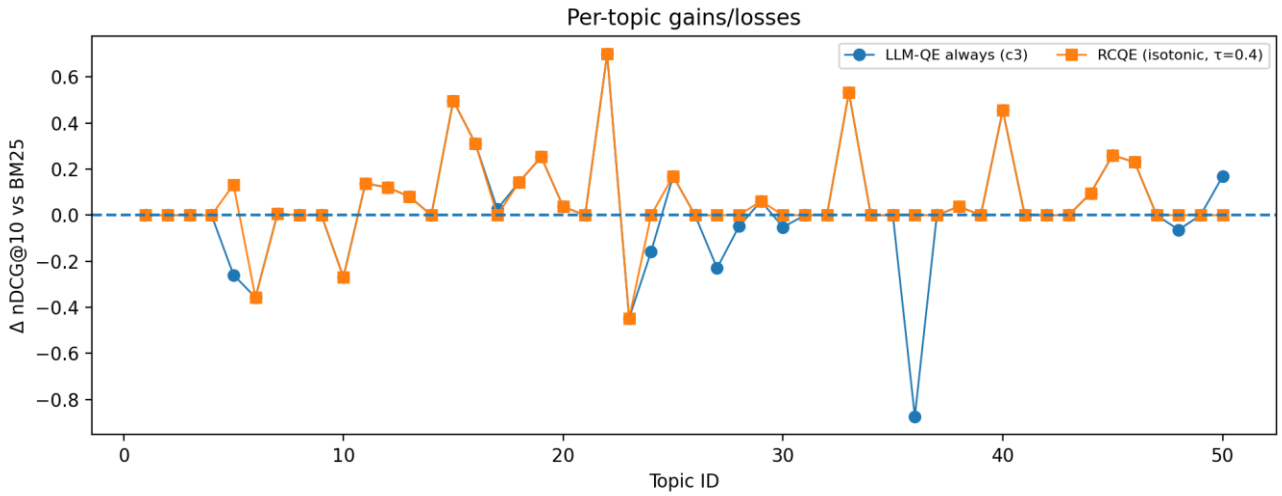

Figure 6. Per-topic ΔnDCG@10 vs. BM25 (Topic IDs 1–50).

Case study: catastrophic drift on Topic 36 (spike structure). Topic 36 is a highly technical virology query. The naive c3 expansion adds a mixture of structural biology terms and general COVID-19 terms, which inadvertently increases matches to broad COVID-19 review articles and epidemiology papers. As a result, its top-ranked documents are irrelevant (Table 13 shows zero relevance for the top 5). In contrast, BM25 already retrieves highly relevant structural papers at the top ranks, and RCQE preserves this ranking by abstaining (or selecting the unexpanded query). This example demonstrates the practical meaning of drift: the expansion changed the top results from precise structural papers to generic COVID-19 discussions, which would be a visible and severe failure to a user.

Table 13. Topic 36 case study: top-5 retrieved documents under BM25, naive expansion (c3), and RCQE (τ=0.4), with relevance grades.

System	Rank	DocID	Rel	Title
BM25	1	alrbutoy	2	Genetic variants in TMPRSS2 and Structure of SARS-CoV-2 spike glycopro…
BM25	2	34ljq0qt	2	Structural Basis of SARS-CoV-2 Spike Protein Priming by TMPRSS2
BM25	3	nq16jcs9	1	A glycan cluster on the SARS-CoV-2 spike ectodomain is recognized by F…
BM25	4	ebbzx8yr	2	A Cryptic Site of Vulnerability on the Receptor Binding Domain of the …
BM25	5	g81ylcxq	2	Structure-based Design of Prefusion-stabilized SARS-CoV-2 Spikes
LLM-QE(c3)	1	noojvz68	0	The epidemiology and clinical characteristics of co-infection of SARS-…
LLM-QE(c3)	2	883t7brb	0	Virology, Epidemiology, Pathogenesis, and Control of COVID-19
LLM-QE(c3)	3	j1sgaido	0	COVID-19: a conundrum to decipher.
LLM-QE(c3)	4	v5pcpss9	0	COVID-19: a conundrum to decipher
LLM-QE(c3)	5	b1cruunn	0	Tracing New Clinical Manifestations in Patients with

				COVID-19 in Chile…
RCQE(τ=0.4)	1	alrbutoy	2	Genetic variants in TMPRSS2 and Structure of SARS-CoV-2 spike glycopro…
RCQE(τ=0.4)	2	34ljq0qt	2	Structural Basis of SARS-CoV-2 Spike Protein Priming by TMPRSS2
RCQE(τ=0.4)	3	nq16jcs9	1	A glycan cluster on the SARS-CoV-2 spike ectodomain is recognized by F…
RCQE(τ=0.4)	4	ebbzx8yr	2	A Cryptic Site of Vulnerability on the Receptor Binding Domain of the …
RCQE(τ=0.4)	5	g81ylcxq	2	Structure-based Design of Prefusion-stabilized SARS-CoV-2 Spikes

Top improvements and harms. Tables 8 and 9 summarize the largest per-topic gains and losses. Both c3 and RCQE share the same top improvements, indicating that RCQE does not suppress large positive gains. However, RCQE removes the worst outlier of c3 (Topic 36) and reduces the number of harmed topics. The remaining RCQE harms are concentrated in a few topics (e.g., Topics 6, 10, 23), suggesting that additional drift detectors could target these query types specifically.

Table 8. Top-5 topic-level improvements (ΔnDCG@10) for naive expansion (c3) and RCQE.

Rank	**Topic (c3)**	**ΔnDCG@10 (c3)**	**Topic (RCQE)**	**ΔnDCG@10 (RCQE)**
1	22	0.700	22	0.700
2	33	0.533	33	0.533
3	15	0.497	15	0.497
4	40	0.455	40	0.455
5	16	0.312	16	0.312

Table 9. Largest topic-level harms (ΔnDCG@10) for naive expansion (c3) and RCQE.

Rank	Topic harmed (c3)	ΔnDCG@10 (c3)	Topic harmed (RCQE)	ΔnDCG@10 (RCQE)
1	36	-0.874	23	-0.447
2	23	-0.447	6	-0.357
3	6	-0.357	10	-0.270
4	10	-0.270		
5	5	-0.259		

Cross-validation stability. Table 6 reports fold-level results for RCQE (isotonic, τ=0.4). Because TREC-COVID contains only 50 topics, fold-to-fold variability is expected. Nevertheless, RCQE maintains higher mean effectiveness than BM25 in most folds and consistently expands a minority of topics, indicating that the abstention mechanism generalizes beyond a single split.

Table 6. RCQE (isotonic, τ=0.4) performance by cross-validation fold.

Fold	nDCG@10	MAP	Coverage
1	0.646	0.192	0.500
2	0.628	0.199	0.600
3	0.539	0.204	0.400
4	0.686	0.283	0.300
5	0.565	0.188	0.500
Mean	0.613	0.213	0.460

Ablations. Table 10 evaluates which signals drive RCQE. Using only lexical drift features reduces nDCG@10 to 0.595 and increases harm probability. Adding retrieval-shape features yields the best overall trade-off, suggesting that ranking structure contains useful drift signals beyond surface overlap. This is encouraging for deployment, because retrieval-shape features can be computed from shallow candidate retrieval without requiring deep semantic modeling.

Table 10. Ablation study for RCQE variants (5-fold CV, τ=0.4).

Variant	nDCG@10	MAP	Coverage	Risk P(harm|exp)	Risk magnitude
RCQE lexical-only + isotonic (τ=0.4)	0.595	0.214	0.500	0.280	0.248
RCQE full features + isotonic (τ=0.4)	0.613	0.213	0.460	0.130	0.358
RCQE full features	0.611	0.211	0.500	0.200	0.232

uncalibrated (τ=0.4)					

Efficiency. Table 11 reports average retrieval time per topic. In our sparse Python BM25 implementation, a single query costs 5.9 ms on average, while the RCQE pipeline costs 57 ms due to evaluating multiple candidates. This cost is acceptable for offline evaluation and can be reduced in practice by restricting candidate count K, computing retrieval-shape features using shallow retrieval (top 10–50), and reusing partial computations across candidates. Because candidate evaluation is parallelizable, practical latency can be reduced further through batching.

Table 11. Average retrieval time per topic in our implementation (Python + sparse BM25).

Method	**Avg seconds/query**
BM25 (single query)	0.0059
LLM-QE (c3, single expanded query)	0.0094
`RCQE pipeline (≈6 BM25 calls/topic)`	0.0574

Practical guidance for setting τ. Coverage–risk curves provide a direct procedure. Given a tolerated harm probability α, one can select the smallest τ such that $P\_harm \leq \alpha$ on a validation set, then deploy with that τ. In our experiments, $\tau \approx 0.6$ achieves $P\_harm \approx 0.10$ at 24% coverage, while $\tau \approx 0.4$ achieves higher effectiveness at 46% coverage. In safety-critical biomedical settings, the conservative regime is attractive because it yields improvements on a subset of topics while largely avoiding drift. In exploratory analysis scenarios, a lower τ may be appropriate to maximize recall and accept occasional drift.

Interpreting drift signals. A practical advantage of RCQE is that its decision features are interpretable and correspond to intuitive drift mechanisms. Table 15 compares feature means across candidates that improve, harm, or produce no gain. Improving candidates add moderately more terms (mean len add=15.9) and, crucially, those added terms have higher idf (add idf mean=5.48), indicating specificity. Harmful candidates add a similar number of terms (len_add=14.0) but have much higher fractions of generic risk terms (risk term frac=0.153) and substantially lower top-10 overlap with the baseline (0.407). This supports the hypothesis that drift is reflected both in the vocabulary profile (generic vs. specific additions) and in ranking behavior (low overlap indicates a sharp change in retrieved evidence). No-gain candidates have the highest overlap (0.539), suggesting that many expansions are "benign but redundant": they largely preserve the retrieved set while not changing top ranks enough to move nDCG@10.

These patterns also clarify why retrieval-shape features helped in the ablation (Table 10). Lexical overlap alone cannot distinguish between two expansions that add similar words but retrieve different documents, especially in a domain like COVID-19 where many terms co-occur broadly. In contrast, top-10 overlap directly measures whether the evidence set changes. In a deployment, a simple rule such as "reject any candidate with overlap below ε" could serve as an additional hard safety guardrail. RCQE's calibrated probability and overlap feature provide complementary signals: calibration controls average harm probability, while overlap can bound worst-case drift by preventing extreme distribution shifts in retrieved documents.

Where RCQE still fails. Although RCQE reduces harm frequency, its remaining errors concentrate in topics whose relevant evidence is split across multiple sub-areas (e.g., comorbidity plus outcomes). In these topics, expansions can change emphasis: adding comorbidity descriptors may shift retrieval toward mechanistic papers, while the information need may be outcome-focused (mortality, severity, prognosis). Better candidate generation (e.g., producing alternatives that explicitly represent different facets of the query) could help here. Another option is to extend the selector to consider multiple objectives, such as constraining changes in both nDCG@10 and recall, or to use conformal risk controls to guarantee that harmful expansions occur below a target rate under exchangeability assumptions [14].

Implications for LLM-based biomedical search. RCQE does not depend on how candidates are generated, only on having a small candidate set and features. In an LLM setting, one can generate candidates by prompting for (i) synonyms and abbreviations, (ii) MeSH descriptors [23], (iii) drug code names, and (iv) paraphrases at different specificity levels, then apply RCQE to decide which ones to use. The coverage–risk curve becomes a

governance tool: stakeholders can select τ based on an acceptable harm rate, validate it on held-out queries, and monitor calibration drift over time. This approach complements content filtering and citation-aware summarization because it operates at the retrieval layer, reducing the chance that downstream LLMs are fed irrelevant evidence due to query drift.

Table 15. Mean feature values for expansion candidates grouped by outcome (all candidates, j>0).

label	len_add	add_idf_mean	risk_term_frac	top10_overlap	top10_entropy
Harm	7.330	3.057	0.332	0.079	0.787
Improve	6.375	4.131	0.033	0.115	0.817
NoGain	4.286	2.991	0.104	0.319	0.840

Table 15 provides an aggregate view of which signals distinguish helpful expansions from harmful ones in our candidate set. These aggregates are computed over all evaluated expansion candidates (excluding the unexpanded query).

Selecting τ in practice also depends on the operational definition of harm. In this paper, harm is defined as ΔnDCG@10<0, which is an early-precision criterion. A biomedical triage system might instead prioritize high-depth recall for systematic review tasks, or might treat any shift in the top results as risky regardless of nDCG. RCQE supports these variants: the label and risk definition can be changed to match the target requirement, and calibration can be repeated for the new notion of correctness. For example, one can train the classifier on a binary label for "recall@100 improves" and then operate RCQE under a recall-based coverage–risk curve. This flexibility is important because biomedical search spans multiple tasks (fact finding, clinical decision support, evidence synthesis) with different failure costs.

Finally, RCQE can be combined with hybrid retrieval architectures. Many biomedical systems use a lexical first stage for recall and a neural reranker for precision. In such systems, expansion should be judged not only by first-stage metrics but also by its effect on the candidate pool presented to the reranker. RCQE can operate at the first stage by predicting whether expansion increases the probability that at least one highly relevant document appears in the top-N candidate set; the selector can then expand only when this probability is high. Alternatively, RCQE can be applied at the query-rewriting layer for dense retrieval, using embedding-based drift features (e.g., cosine similarity between original and rewritten query vectors) and calibrating improvement probabilities with the same isotonic or sigmoid techniques [9]–[12]. This suggests that risk-calibrated selective expansion is a general pattern that can be applied across retrieval backbones as long as an evaluation signal is available.

Monitoring and maintenance are essential for any calibrated decision rule. Calibration can drift when the corpus changes (new papers, new terminology) or when user queries shift. In a deployed biomedical search engine, RCQE therefore benefits from periodic recalibration using newly judged data, or from online evaluation proxies such as inter-annotator agreement and click-based satisfaction signals. Because isotonic regression is lightweight, recalibration is computationally cheap once a small calibration set is available. Operationally, the system can log predicted probabilities, decisions (expand/abstain), and downstream quality indicators, then periodically recompute reliability curves similar to Figure 3. If the observed empirical improvement rates in high-confidence bins fall below the predicted rates, τ can be increased to restore safety until recalibration is performed. This "calibrate, threshold, monitor" cycle turns uncertainty calibration into an explicit governance mechanism for robust biomedical retrieval.

Limitations

First, our candidate generator is intentionally lightweight and deterministic to ensure full reproducibility without external APIs. Although candidates mimic common LLM outputs (synonyms, abbreviations, MeSH-style descriptors [23], and drug code names), stronger generative models could produce richer candidates and may change the absolute performance numbers. The oracle results in Table 3 suggest that better candidates and better gain predictors are both valuable.

Second, our selector uses topic-level supervision derived from the same test collection. While we used 5-fold cross-validation and out-of-fold calibration, TREC-COVID contains only 50 topics, which limits the stability of learned thresholds and calibration curves. Larger biomedical test sets, or training selectors on external collections and evaluating on TREC-COVID in a transfer setting, would more closely match real deployments.

Third, we evaluated only lexical BM25 retrieval. BEIR includes neural baselines and dense retrievers [1], and recent LLM-based retrieval augmentation methods apply to dense retrieval as well [17], [18]. Extending RCQE to dense or hybrid retrieval requires redefining drift features (e.g., embedding-space dispersion) but the same calibrated selective decision applies.

Finally, our risk definitions focus on nDCG@10 harm. Biomedical search systems may require additional risk criteria, such as recall at high depth, safety filters, or per-topic worst-case constraints. Integrating conformal prediction-style guarantees [14] or cost-sensitive rejection rules [13] is a promising direction.

Additionally, our results are specific to a lexical BM25 backbone and to the candidate set we constructed. Stronger neural retrievers can change the effect of lexical expansion; prior studies show that expansions may help weaker retrievers but harm strong rerankers [19]. RCQE's selective mechanism remains applicable, but the feature set would need adaptation (e.g., embedding-space drift indicators).

Finally, the evaluation uses offline relevance judgments. In live biomedical search, user satisfaction and safety outcomes depend on presentation, filtering, and clinical context. RCQE should therefore be validated with user studies and with task-specific safety metrics before deployment.

Conclusion

Risk-Calibrated Query Expansion (RCQE) treats biomedical expansion as a selective decision under uncertainty. On BEIR TREC-COVID, naive always-expand strategies improved average metrics but produced heavy-tailed failures. RCQE combined multi-candidate expansion with calibrated uncertainty to control drift, improving nDCG@10 from 0.549 (BM25) to 0.613 under 5-fold cross-validation while lowering the conditional probability of harm among expanded topics. Coverage–risk curves provided an interpretable knob to trade off safety and effectiveness. These findings support a practical deployment recipe for biomedical search: generate multiple plausible expansions, score them, calibrate uncertainty, and expand only when the predicted gain is reliable.

References

[1] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," arXiv:2104.08663, 2021.

[2] K. Roberts, T. Alam, S. Bedrick, D. Demner-Fushman, K. Lo, I. Soboroff, E. M. Voorhees, L. L. Wang, and W. R. Hersh, "Searching for Scientific Evidence in a Pandemic: An Overview of TREC-COVID," J. Biomed. Inform., vol. 121, p. 103865, 2021.

[3] E. M. Voorhees, K. Roberts, I. Soboroff, and W. R. Hersh, "TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection," SIGIR Forum, vol. 54, no. 1, 2021, doi: 10.1145/3451964.3451965.

[4] C. Carpineto and G. Romano, "A Survey of Automatic Query Expansion in Information Retrieval," ACM Comput. Surv., vol. 44, no. 1, pp. 1–50, 2012, doi: 10.1145/2071389.2071390.

[5] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[6] V. Lavrenko and W. B. Croft, "Relevance-Based Language Models," in Proc. 24th ACM Int. Conf. Research and Development in Information Retrieval (SIGIR), 2001, pp. 120–127, doi: 10.1145/383952.383972.

[7] C. Zhai and J. Lafferty, "Model-Based Feedback in the Language Modeling Approach to Information Retrieval," in Proc. 10th ACM Int. Conf. Information and Knowledge Management (CIKM), 2001, pp. 403–410.

[8] K. Järvelin and J. Kekäläinen, "Cumulated Gain-Based Evaluation of IR Techniques," ACM Trans. Inf. Syst., vol. 20, no. 4, pp. 422–446, 2002, doi: 10.1145/582415.582418.

[9] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. 34th Int. Conf. Machine Learning (ICML), 2017, pp. 1321–1330.

[10] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in Advances in Large Margin Classifiers. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.

[11] B. Zadrozny and C. Elkan, "Transforming Classifier Scores into Accurate Multiclass Probability Estimates," in Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2002, pp. 694–699, doi: 10.1145/775047.775151.

[12] A. Niculescu-Mizil and R. Caruana, "Predicting Good Probabilities with Supervised Learning," in Proc. 22nd Int. Conf. Machine Learning (ICML), 2005, pp. 625–632, doi: 10.1145/1102351.1102430.

[13] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.

[14] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. New York, NY, USA: Springer, 2005.

[15] R. Nogueira, W. Yang, J. Lin, and K. Cho, "Document Expansion by Query Prediction," arXiv:1904.08375, 2019.

[16] R. Nogueira and J. Lin, "From doc2query to docTTTTTquery," Online preprint, 2019.

[17] L. Bonifacio, I. Abonizio, V. Jeronymo, R. Pereira, and R. Lotufo, "InPars: Data Augmentation for Information Retrieval using Large Language Models," arXiv:2202.05144, 2022.

[18] L. Gao, Z. Dai, and J. Callan, "Precise Zero-Shot Dense Retrieval without Relevance Labels (HyDE)," arXiv:2212.10496, 2022.

[19] O. Weller, K. Lo, D. Wadden, D. Lawrie, B. Van Durme, A. Cohan, and L. Soldaini, "When Do Generative Query and Document Expansions Fail? A Comprehensive Study Across Methods, Retrievers, and Datasets," arXiv:2309.08541, 2023.

[20] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A Note on Platt's Probabilistic Outputs for Support Vector Machines," Mach. Learn., vol. 68, no. 3, pp. 267–276, 2007, doi: 10.1007/s10994-007-5018-6.

[21] R. Jagerman, H. Zhuang, Z. Qin, X. Wang, and M. Bendersky, "Query Expansion by Prompting Large Language Models," arXiv:2305.03653, 2023.

[22] L. L. Wang et al., "CORD-19: The COVID-19 Open Research Dataset," arXiv:2004.10706, 2020.

[23] U.S. National Library of Medicine, "Medical Subject Headings (MeSH)," 2023. [Online]. Available: https://www.nlm.nih.gov/mesh/.

[24] Hanqi Zhang, "DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models", JACS, vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[25] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, "Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)", JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.

[26] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, "ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence", JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[27] Hanqi Zhang, "Risk-Aware Budget-Constrained Auto-Bidding under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework", JACS, vol. 4, no. 6, pp. 30–47, Jun. 2024, doi: 10.69987/JACS.2024.40603.

[28] T. Shirakawa, Y. Li, Y. Wu, S. Qiu, Y. Li, M. Zhao, H. Iso, and M. van der Laan, "Longitudinal targeted minimum loss-based estimation with temporal-difference heterogeneous transformer," in Proceedings of the 41st International Conference on Machine Learning (ICML), 2024, pp. 45097–45113, Art. no. 1836.

[29] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," arXiv preprint arXiv:2408.05944, 2024.

[30] Z. S. Zhong and S. Ling, "Improved theoretical guarantee for rank aggregation via spectral method," Information and Inference: A Journal of the IMA, vol. 13, no. 3, 2024.

[31] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," Proceedings of the 6th International Conference on Computing and Data Science (ICCDS), 2024.

[32] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024.

[33] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024, pp. 89–94.

[34] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFAConv and triplet attention," Proceedings of the 2nd International Conference on Software Engineering and Machine Learning (SEML 2024), 2024.

[35] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, "Implementation of an AI-based MRD Evaluation and Prediction Model for Multiple Myeloma", FCIS, vol. 6, no. 3, pp. 127–131, Jan. 2024, doi: 10.54097/zJ4MnbWW.