BEYOND THE QUERY: CORPUS-CONDITIONED PREDICTIVE RETRIEVAL FOR UNDERSPECIFIED QUERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Users search because they may lack the knowledge needed to describe the information they seek. Query expansion can help, but may emphasize only some concepts or relations relevant to an underspecified query. We introduce CPR (Corpus-conditioned Predictive Retrieval), which uses corpus documents as partial representations of possible intents. Motivated by the posterior-predictive principle, CPR averages corpus-based relevance predictions under a support prior constructed from semantic and lexical evidence to refine candidate support, without additional training. Experiments on reasoning-intensive retrieval show that better prior rankings need not yield better document predictions. Composition with query expansion recovers shared failures, while weaker direct retrievers can still provide useful document relations. These findings support evaluating evidence through the predictions it enables and using corpus evidence to complement generated queries and stronger embeddings.

1 INTRODUCTION

Dense retrieval has become a standard approach to information retrieval, encoding queries and documents in a shared embedding space and ranking them by query–document similarity (Zhao et al., 2024; Lee et al., 2025). Documents are therefore preferred according to how closely they match what the query expresses. Users, however, often search precisely because they lack the knowledge needed to describe the desired information. The problem is especially consequential in reasoning-intensive retrieval, where relevance may depend on a mechanism, concept, or distinction that the query does not name (Su et al., 2025; Zhang et al., 2025a). A query can describe a problem with discriminative terms while leaving concepts or relations needed to identify relevant documents unspecified. Its observed terms provide an anchor, but not a complete criterion of relevance. This tension motivates our central question.

Can an underspecified query alone determine the criterion of relevance?

Prior work on query augmentation uses large language model (LLM)-generated documents and passages (Gao et al., 2023; Wang et al., 2023a), corpus-guided expansion and reasoning (Lei et al., 2024; 2025), or query rewriting and decomposition (Qin et al., 2025; Liu et al., 2025). Corpus-guided retrieval uses retrieved documents for query refinement or embedding-dimension selection (Yu et al., 2021; Wang et al., 2023b; D'Erasmo et al., 2025; 2026), and exploits document neighborhoods, semantic concepts, or distinctions between documents (Kulkarni et al., 2024; Kang et al., 2024; Zhang et al., 2025c; Ashok et al., 2026). These approaches enrich retrieval beyond the query. However, for an underspecified query, a generated expansion may emphasize only some relevant concepts or relations. Relying solely on this evidence can narrow the relevance criterion.

Recent work considers evidence aggregation and probabilistic retrieval. QUARK (Lyu et al., 2026) uses LLM-generated recovery hypotheses, while QuDAR (Kim et al., 2026) combines original and expanded evidence. PQR (Kang et al., 2025) models LLM-generated potential-query embeddings as document-specific distributions, whereas UEC (Lim et al., 2026) uses supervised post-hoc fitting for embedding ensembling. These approaches represent alternatives through generation or additional modeling. Retrieval, however, already has access to the corpus from which relevant documents must be identified. Its documents may contain terminology, concepts, or relations left unnamed by the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

query, providing partial representations of possible intents. We therefore use the corpus not only as a candidate collection, but also as evidence for relevance prediction.

To formalize this, let q and C denote the observed query and corpus, and let z represent aspects of the information need left unspecified by q. Predicting relevance across possible intents motivates the posterior-predictive decomposition (Murphy, 2022),

$$p(R_d = 1 \mid q, C) = \int p(R_d = 1 \mid d, q, z, C)\, p(z \mid q, C)\, \mathrm{d}z. \tag{1}$$

The indicator R_d denotes whether document d is relevant. Applying the decomposition requires defining *possible intents*, their *support prior*, and the *relevance likelihood* under each intent. We introduce **CPR** (Corpus-conditioned Predictive Retrieval), which constructs these components from corpus evidence without additional training. We use **corpus documents as support points** for possible intents, retaining eligible sources beyond the initial query neighborhood. The **support prior** $p(z \mid q, C)$ equally combines semantic and semantic–lexical priors, incorporating query terms while retaining semantic support when useful terminology is unnamed. The **relevance likelihood** $p(R_d = 1 \mid d, q, z, C)$ uses semantic relations relative to the corpus background and lexical coverage. We average these predictions under the support prior to refine each candidate's query support.

We evaluate CPR on 11 BRIGHT tasks (Su et al., 2025) and eight R2MED tasks (Zhang et al., 2025a) using three frozen retrievers. Across tasks and backbones, CPR raises mean nDCG@10 from 20.30 to 27.04 and Recall@100 from 50.60 to 63.92 over direct retrieval. Our analyses show that a prior's standalone ranking need not reflect its contribution to prediction. Sources beyond the initial neighborhood can improve candidate evaluation, and CPR's gains tend to be larger when query terms distinguish relevant documents from competitors. Composition with query expansion recovers shared failures, while weaker direct retrievers retain useful document relations. These findings support better use of corpus evidence as complementary to generated queries and stronger embeddings.

2 Corpus-conditioned Predictive Retrieval

Our goal is to identify relevant documents from the query q using the corpus $C = \{d_1, d_2, \cdots, d_N\}$ itself[1]. We now define the latent intent z in Eq 1 using the available corpus. The mixture in Eq 1 allows us to use multiple intent candidates to represent a query. We assume that a document $z \in C$ might partially represent the user's hidden intent when its content is supported by the query[2]. Thus,

$$p(R_d|q) = \sum_{z \in \mathcal{A}_q} p(R_d|d, q, z) p(z|q), \tag{2}$$

where $\mathcal{A}_q \subseteq C$ contains all corpus documents except those excluded for q by the benchmark protocol, with no top-k restriction. Keeping all eligible documents available allows sources beyond the initial query neighborhood to contribute.

To compute this equation, we need two components: *relevance likelihood* $p(R_d|d, z, q)$ and *support prior* $p(z|q)$. The support determines how strongly each document contributes as an intent candidate, while the likelihood uses that candidate to predict relevance. Eq 2 specifies how predictions are averaged, but there are many possible methods for estimating the two components. We use simple estimators here, and these components can be extended with more complex ones.

2.1 Support Prior

Let us compute the support prior $p(z|q)$ in Eq 2 first. We estimate the prior as an unweighted mean of two probabilistic distributions π_s and π_{sl}:

$$p(z|q) = \frac{\pi_s(z|q) + \pi_{sl}(z|q)}{2}. \tag{3}$$

Here, π_s provides semantic support, while π_{sl} evaluates both lexical overlap and semantic compatibility. Lexical matches thus contribute in conjunction with meaning, rather than as an independent

[1] For simplicity, we will omit the common random variable C from here.

[2] As we consider a retrieved document as a signal of intent, we override the notation z to represent a document in the corpus.

basis for intent support. We retain π_s because the query might be underspecified; relevant document may not share same lexicon with the query.

Semantic prior π_s. All cosine similarities below use embeddings that are normalized, centered by the mean normalized corpus vector, and normalized again. To distinguish affinity between q and z from similarity shared nearby z in C, we apply a CSLS-inspired density correction (Conneau et al., 2017):

$$\mathrm{sim}^{(s)}(q, z) = 2\cos(q, z) - \bar{\sigma}(z), \tag{4}$$

where $\bar{\sigma}(z)$ averages $\cos(z, x)$ over the k nearest non-self neighbors of z, with $k = \lceil\sqrt{N}\rceil$.

Though $\mathrm{sim}^{(s)}$ adjusts for neighborhood density, we use its relative ordering rather than score magnitudes or gaps to assign intent support. Inspired by power laws, we adopt reciprocal ranks to give more weight to documents with high $\mathrm{sim}^{(s)}$ while retaining support for lower-ranked documents. Let $\rho_s[z|q]$ denote the descending rank of document z within $\mathcal{A}_q$, based on $\mathrm{sim}^{(s)}(q, z)$. Then, we compute π_s as a finite reciprocal distribution over $\rho_s[z|q]$:

$$\pi_s(z|q) = \mathrm{Reciprocal}(\rho_s[z|q]), \qquad \mathrm{Reciprocal}(x) \propto x^{-1} \quad \forall x \in \{1, 2, \ldots, |\mathcal{A}_q|\}, \tag{5}$$

where $\mathrm{Reciprocal}(x)$ represents a finite reciprocal distribution.

Semantic-lexical prior π_{sl}. Semantic evidence itself may not capture lexical similarity between q and z, and vice versa. So, we design semantic-lexical evidence π_{sl}. To weigh more on overlap between important lexicons, we sum up inverse document frequency $\mathrm{IDF}(t)$ for common terms:

$$\mathrm{sim}^{(l)}(q, z) = \sum_{t \in q \cap z} \mathrm{IDF}(t). \tag{6}$$

Using $\mathrm{sim}^{(l)}$, we define $\rho_l[z|q]$ as the descending lexical rank, assigning tied scores as their average rank. To compute π_{sl}, we combine the two ranks: We re-rank the documents using the product $\rho_s[z|q] \times \rho_l[z|q]$ in ascending order. Multiplication treats proportional improvements in either rank equally. As a result, we obtain combined rank $\rho[z|q]$ and the semantic-lexical prior:

$$\pi_{sl}(z|q) = \mathrm{Reciprocal}(\rho[z|q]). \tag{7}$$

2.2 Relevance Likelihood

Now, let us compute the relevance likelihood $p(R_d|d, z, q)$ in Eq 2. As R_d is unobservable during the search, we further model the relevance as the relationship between a document d and a source document z, which supports a latent intent:

$$p(R_d = 1|d, z, q) \propto \frac{r_s(d|z) r_l(d|q)}{u(d)}, \qquad u(d) = \frac{1}{N} \sum_{z' \in C} r_s(d|z'), \tag{8}$$

where the binary variable R_d denotes relevance, and $r_s(d|z)$ and $r_l(d|q)$ predict semantic and lexical relevance scores of document d, respectively. Here, we use ratio between $r_s r_l$ and u to represent how much z favors d compared to its average support documents; thus, we set $u(d)$ as r_s marginalized over latent intents in C. For the details, see Appendix A.3.

Semantic relevance score r_s. This function is designed to capture relevance using the similarity between a document d and a latent intent z. Here, we compute $\rho_s[d|z]$ as the descending rank of d among documents in $C \setminus \{z\}$ by $\cos(z, d)$. We compute r_s as:

$$r_s(d|z) \propto (\rho_s[d|z])^{-1} \quad \forall d \neq z; \qquad r_s(z|z) = 0, \tag{9}$$

with $\sum_{d \in C} r_s(d|z) = 1$. We retain the leading rank masses and distribute the remaining mass uniformly over the other non-self documents (Appendix A.1). Thus, r_s gives greater support to leading neighbors while retaining mass on the remaining non-self documents.

Lexical relevance score r_l. To consider both query terms and source-supported terminology, we compute r_l with two lexical factors, $r_l(d|q) \propto \lambda_q(d)\lambda_{\text{src}}(d)$, with $\sum_d r_l(d|q) = 1$. Here, $\lambda_q(d)$ and $\lambda_{\text{src}}(d)$ measure query-term and source-supported terminology coverage, respectively. Each factor averages one with the ratio of the candidate's weighted coverage to its background expectation. Query weights w_q emphasize terms occurring less frequently in source documents, while source weights w_{src} favor non-query terms whose occurrence increases under $p(z|q)$ relative to uniform source weights. We estimate the weights with other corpus documents (Appendix A.2).

2.3 Document Scoring and Selection

The estimation for Eq 2 discussed above assumes that the intent z is not exactly equal to our retrieved document d. But, in the real world, d could be both a latent intent and the target retrieved document. So, we need to consider two probabilities at once: (1) the probability of selecting d directly from q and (2) the probability of selecting d indirectly from q. By the definitions, the first probability is $p(z = d|q)$ and the second probability is $p(R_d|q)$. Thus, we use the following probability for the document scoring:

$$g(d) = p(z = d|q) \cdot p(R_d|q) = p(z = d|q) \cdot \sum_{z \in \mathcal{A}_q \setminus \{d\}} \frac{r_s(d|z) r_l(d|q)}{u(d)} p(z|q). \tag{10}$$

Within the summation, semantic relations are averaged over sources and combined with lexical coverage. The outer factor weights this prediction by the candidate's own query support. The common proportionality constant in Eq 8 is omitted because it does not affect document ordering.

Finally, we preserve the document with the highest $p(z = d|q)$ as the first result and rank the remaining eligible documents by $g(d)$. Ties are resolved by corpus row.

3 Experiment Setup

Benchmarks. We evaluate BRIGHT (Su et al., 2025) and R2MED (Zhang et al., 2025a) to examine whether corpus evidence helps reasoning-intensive retrieval beyond direct query matching. The 19 tasks span science, social science, technology, mathematics, and biomedicine, comprising 11 BRIGHT and eight R2MED tasks with shared Biology counted only under R2MED (Appendix B).
Retrievers. We use three frozen open-source retrievers. E5-large-v2 (Wang et al., 2022) for general-purpose retrieval, Qwen3-Embedding-4B (Zhang et al., 2025b) for LLM-based embeddings, and ReasonIR-8B (Shao et al., 2025), trained for reasoning-intensive retrieval.

Baselines. We compare three strategies for supplementing query evidence: *query augmentation*, *corpus-guided retrieval*, and *evidence aggregation and probabilistic retrieval*. Base, BM25, and their rank product (RP) provide direct-matching references. HyDE (Gao et al., 2023) and ThinkQE (Lei et al., 2025) use hypothetical-document generation and iterative expansion, respectively. Corpus-guided methods include ECLIPSE-PRF and RDIME-PRF (D'Erasmo et al., 2025; 2026) for feedback-based dimension selection, LexBoost (Kulkarni et al., 2024) for document-neighborhood scoring, and SemRank (Zhang et al., 2025c) for LLM-selected corpus concepts. The third group includes document-prior scoring (Parry et al., 2026) and QuDAR/QuDAR-LLM (Kim et al., 2026). Both QuDAR variants use generated passages, combining original and expanded sparse–dense evidence through score-based confidence and LLM judgments.

Methods share cached original-query and document embeddings. LLM components use Qwen3-8B (Yang et al., 2025) with official prompts and retrieval procedures, reusing generated evidence across retrievers. CPR+HyDE holds generated evidence fixed, using standalone HyDE's documents and semantic input, with original-query and generated terms for lexical evidence (Appendix C).

Evaluation. We report nDCG@10 for ranking quality and Recall@50/100 for coverage, averaging queries within each task–backbone cell and then weighting tasks and backbones equally.

4 Results

Table 1 reports nDCG@10 across 19 tasks and three retrievers, yielding three findings:

Table 1: nDCG@10 on 19 tasks; Avg. is their unweighted mean. Bold/underline mark best/second-best separately in the LLM-free and LLM-assisted (†) groups of each backbone, shown above/below the divider. RP: dense–BM25 rank product.

	BRIGHT											R2MED								
Method	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.	Bio.	Bioinf.	MedSci.	MedX.	MedQA	PMC-T	PMC-C	IIYi	**Avg.**
Lexical																				
BM25	28.08	16.45	13.42	10.91	16.18	16.13	24.71	4.34	6.45	7.33	2.14	18.40	19.81	17.51	1.41	1.54	17.86	20.94	10.62	13.38
E5-large-v2																				
Base	36.96	8.68	8.32	12.17	21.31	9.62	19.89	5.73	5.27	11.46	4.80	13.02	20.38	28.38	2.72	3.71	8.19	14.64	4.90	12.64
RP	**46.40**	20.01	20.33	15.78	**25.01**	16.67	23.97	5.63	**8.26**	11.73	4.64	22.96	24.45	31.47	1.97	3.65	16.17	21.44	9.72	17.38
LexBoost	27.26	15.44	13.10	10.52	18.06	16.15	23.43	4.92	5.27	6.20	1.57	16.93	19.80	18.57	1.18	2.51	**17.88**	21.54	10.63	13.21
ECLIPSE-PRF	37.84	12.07	9.61	11.54	17.18	13.57	19.40	6.80	3.20	11.96	3.44	13.94	20.90	29.04	**3.92**	4.02	8.00	13.30	6.14	12.94
RDIME-PRF	37.45	6.26	6.44	11.24	23.16	9.29	18.85	4.80	4.22	11.74	4.28	13.20	19.82	29.62	3.92	3.47	7.81	12.98	5.51	12.32
Parry et al. (2026)	37.81	10.10	8.62	13.35	21.04	11.46	18.20	8.63	5.30	11.70	3.64	17.84	24.66	29.17	2.72	3.74	8.30	9.06	3.51	13.10
CPR	42.11	**23.74**	**27.16**	**19.20**	22.63	**21.33**	**25.93**	**21.17**	5.87	**17.06**	**9.05**	**38.06**	**34.58**	**34.67**	3.24	**5.70**	17.07	**26.69**	**11.38**	**21.40**
HyDE†	45.31	16.47	16.09	15.30	**29.91**	14.52	17.39	8.48	5.18	16.38	11.80	36.66	31.73	39.26	**17.64**	**16.59**	23.52	17.51	9.23	20.47
SemRank†	26.79	13.23	12.42	10.86	13.42	11.00	**23.78**	5.39	5.02	12.16	5.25	3.73	23.06	29.90	3.06	5.36	11.43	18.02	12.08	12.94
ThinkQE†	46.38	15.54	22.59	15.01	27.04	20.91	18.04	9.98	6.80	17.18	9.68	37.58	33.51	42.66	5.88	10.17	16.24	19.95	12.21	20.39
QuDAR†	44.64	16.04	13.51	14.49	25.91	13.49	20.25	8.85	6.72	12.31	5.42	24.87	31.43	36.71	4.54	6.31	11.61	17.74	8.50	17.02
QuDAR-LLM†	45.52	17.09	17.38	15.17	26.33	16.88	20.52	8.72	7.22	13.59	6.65	31.20	34.30	38.75	4.46	5.85	12.91	18.00	8.02	18.35
CPR+HyDE†	**46.96**	**23.92**	**36.54**	**25.98**	28.61	**31.12**	23.27	**35.82**	**9.88**	**23.32**	**14.58**	**55.61**	**43.15**	**44.55**	15.05	16.18	**34.17**	**27.82**	**14.82**	**29.02**
Qwen3-Embedding-4B																				
Base	34.29	15.27	22.67	13.07	17.55	14.62	33.00	1.73	8.53	34.23	**31.98**	15.98	48.30	40.62	13.56	18.97	37.88	34.58	21.05	24.10
RP	37.46	18.83	26.03	17.11	23.36	18.61	34.86	5.70	**11.88**	20.77	17.78	19.52	38.41	35.22	7.25	12.09	35.12	36.54	22.17	23.09
LexBoost	28.42	15.56	13.66	8.99	17.32	15.45	23.68	3.88	5.54	6.22	2.26	15.77	22.00	18.61	1.02	2.53	18.27	23.59	11.15	13.36
ECLIPSE-PRF	33.49	15.55	23.99	13.74	16.87	17.86	32.97	1.45	8.66	**34.79**	31.15	17.14	48.19	39.35	14.31	19.57	41.41	37.27	28.44	25.06
RDIME-PRF	36.25	13.89	22.55	11.86	15.81	14.37	32.01	1.04	8.19	33.09	31.47	16.00	48.43	40.58	14.39	19.40	40.53	38.36	**28.49**	24.56
Parry et al. (2026)	33.14	15.74	23.72	16.39	16.44	19.51	33.00	1.40	9.17	32.53	27.19	21.98	43.75	37.84	**15.48**	**20.53**	**44.09**	30.41	25.77	24.64
CPR	**47.74**	**25.68**	**37.75**	**23.73**	**29.06**	**27.13**	**40.76**	**19.78**	9.64	32.88	25.55	**37.03**	**49.35**	**40.69**	9.76	17.37	43.73	**40.90**	25.61	**30.74**
HyDE†	35.79	19.01	27.12	15.20	17.90	19.76	32.67	2.23	9.38	**39.41**	**37.90**	21.97	52.14	45.48	**21.66**	**24.76**	43.66	34.27	**30.06**	27.91
SemRank†	32.23	14.35	20.00	11.70	14.75	12.42	35.12	1.82	7.72	32.62	31.40	16.59	43.15	40.66	14.35	18.44	38.06	31.72	20.39	23.03
ThinkQE†	29.36	9.12	13.72	9.01	14.29	5.98	35.31	0.92	8.72	38.30	28.69	11.73	45.17	34.53	9.35	13.70	30.18	33.23	24.38	20.82
QuDAR†	34.22	15.49	24.98	14.20	19.42	16.58	30.51	1.64	9.05	33.49	30.58	19.91	51.98	45.28	12.35	20.74	35.71	36.72	27.13	25.26
QuDAR-LLM†	34.17	15.77	25.94	13.97	19.59	16.82	30.25	1.88	8.74	33.86	32.48	24.82	51.69	46.86	12.65	21.61	32.89	36.59	27.33	25.68
CPR+HyDE†	**51.00**	**27.61**	**42.72**	**28.26**	**31.70**	**33.33**	**39.00**	**31.09**	**9.93**	34.52	36.28	**44.52**	**56.85**	**49.75**	17.09	22.69	**44.55**	**36.72**	29.05	**35.09**
ReasonIR-8B																				
Base	31.34	23.52	30.17	18.02	23.74	20.27	**34.85**	10.22	14.56	31.67	27.16	19.73	40.12	40.04	9.46	12.95	26.00	26.67	18.51	24.16
RP	35.17	21.52	27.49	17.70	24.54	20.21	32.83	8.27	13.40	18.61	17.16	24.89	34.14	34.65	5.95	10.17	28.40	33.89	20.25	22.59
LexBoost	27.65	15.70	13.65	10.36	18.11	15.72	22.77	4.77	4.07	6.11	2.01	15.40	19.90	19.32	0.99	2.42	18.73	23.00	10.86	13.24
ECLIPSE-PRF	31.27	21.92	28.90	17.60	20.70	21.87	34.69	8.49	**14.96**	**32.03**	27.04	25.29	44.36	39.24	11.02	15.32	**34.92**	30.02	20.31	25.26
RDIME-PRF	31.23	23.60	30.31	17.87	21.65	21.93	33.72	9.90	13.13	30.37	**27.30**	24.99	44.24	**40.29**	**12.97**	15.69	32.25	31.72	21.09	25.49
Parry et al. (2026)	32.11	24.93	31.07	21.86	23.88	22.77	28.23	15.00	14.63	30.26	21.69	31.87	40.94	39.47	11.88	**15.82**	33.38	25.34	18.15	25.44
CPR	**48.72**	**26.89**	**33.60**	**24.35**	**28.31**	**27.35**	33.45	**15.37**	14.84	26.24	21.22	**47.14**	**47.58**	39.82	8.98	13.99	32.20	**37.82**	**22.36**	**28.96**
HyDE†	36.11	28.15	37.63	24.48	29.58	24.41	36.65	15.26	11.63	40.37	**36.99**	35.66	46.42	44.74	**22.22**	**22.76**	**43.33**	30.51	**28.44**	31.33
SemRank†	27.34	16.49	25.37	14.89	18.31	15.47	33.45	14.09	12.39	28.55	20.84	20.93	34.90	41.11	8.85	13.68	24.65	23.80	16.94	21.69
ThinkQE†	33.62	18.35	23.18	14.78	22.20	17.14	34.69	10.44	13.56	**41.27**	27.00	19.97	41.52	36.02	10.79	12.87	32.16	33.09	19.40	24.32
QuDAR†	34.20	23.41	31.82	19.37	27.79	21.74	**37.43**	15.99	12.83	30.39	26.71	35.04	47.77	44.05	13.50	17.81	29.90	32.59	25.03	27.76
QuDAR-LLM†	35.07	23.98	31.38	19.92	28.37	22.30	37.31	15.54	12.89	31.34	27.95	37.73	47.31	43.50	13.40	18.17	29.68	32.88	24.98	28.09
CPR+HyDE†	**52.74**	**29.15**	**40.09**	**32.29**	**32.25**	**36.07**	32.01	**33.66**	**13.80**	31.03	32.25	**56.89**	**51.14**	**44.96**	17.13	20.51	40.47	**34.97**	22.25	**34.40**

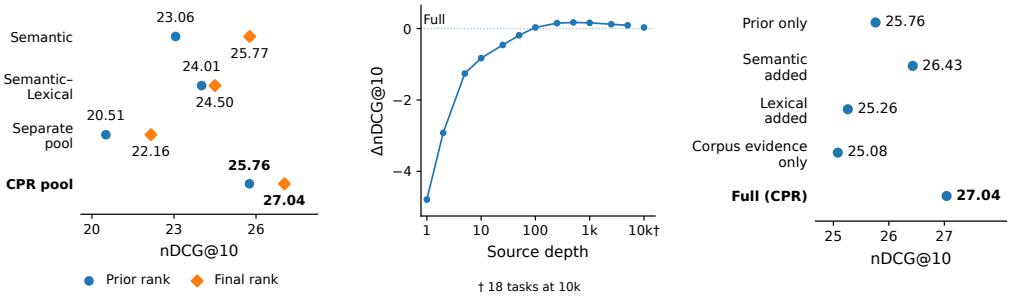
CPR improves ranking quality and relevant-document coverage. CPR exceeds the compared LLM-free methods in task-average nDCG@10 on every backbone and remains competitive with LLM-assisted methods. Gains also extend to Recall@50 and 100. CPR leads the compared standalone methods with E5 and Qwen3, while HyDE retains a recall advantage with ReasonIR (Appendix D). We examine how the support prior, relevance likelihood, and document scoring contribute to these gains (RQ1).

The benefit varies across tasks and query evidence. CPR shows advantages on Biology and Earth Science but trails HyDE on TheoremQA-Theorems and PMC Treatment. Their different task-level strengths motivate examining if generated evidence complements corpus-conditioned prediction. Combining CPR with HyDE improves task-average nDCG@10 over both methods on every backbone. We examine these differences and composition's effects on recovery and ranking (RQ2).

The gain over direct retrieval varies across backbones. CPR yields the largest task-average improvement over Base with E5, followed by Qwen3 and ReasonIR. This raises the question of why E5, despite its weaker direct retrieval, benefits more from corpus-conditioned prediction. We will further examine query alignment and document relations to understand how the evidence available in each backbone and its use in prediction contribute to these differences (**RQ3**).

5 RQ1: How do CPR's components contribute to retrieval?

We examine the support prior, relevance likelihood, and document scoring. Figure 1 summarizes the comparisons; protocols and detailed results appear in Appendix E.



(a) Support prior (b) Relevance likelihood (c) Document scoring

Figure 1: Task–backbone macro nDCG@10. The source-depth curve shows differences from Full.

5.1 Support Prior

Query terms can distinguish possible intents, but useful sources may use different terminology. We compare Semantic (π_s), Semantic–Lexical (π_{sl}), Separate pool ($(\pi_s + \pi_l)/2$), and CPR pool ($(\pi_s + \pi_{sl})/2$), where π_l uses lexical evidence alone. We evaluate each prior's direct ranking and its final ranking after relevance prediction, recomputing prior-dependent evidence.

A better prior ranking need not yield better document predictions. Semantic–Lexical ranks better initially than Semantic, yet Semantic obtains the higher final score. CPR pooling leads at both stages, while Separate pooling falls below both single priors (Figure 1a; Appendix Table 7). This favors retaining semantic-only support alongside semantic–lexical evidence. A prior should therefore be assessed by both its own ranking and the document predictions it enables.

5.2 Relevance Likelihood

An incomplete query may place useful sources outside its initial retrieval neighborhood. To test their contribution, we vary the source pool from Base's top-ranked document to the full eligible corpus, keeping outputs, candidate support, and corpus statistics fixed. Source weights are renormalized to recompute semantic predictions and source-derived terminology.

Prediction benefits from sources beyond the top few. Restricting prediction to a few sources lowers mean nDCG@10, but top-100 approaches Full and top-1,000 slightly improves all three metrics over Full (Figure 1b; Appendix Table 8). After separately normalizing source mass, sources outside the initial top-100 pool favor labeled positives over Base's top-ten competitors not labeled relevant on average, whereas inside sources favor those competitors (Appendix Table 9). RQ2 examines the task differences behind this average.

5.3 Document Scoring and Selection

Related sources may support different aspects of a candidate's content. We compare Prior only, Semantic added, Lexical added, and Full to examine how semantic and lexical evidence refine candidate support. Source weights, retained factors, and the first result stay fixed. Corpus evidence only removes candidate support while retaining query-weighted source predictions.

Lexical evidence strengthens semantic prediction. Semantic evidence improves ranking and recall on its own. Adding lexical evidence to Prior only lowers mean nDCG@10, but the same factor improves nDCG@10 and Recall@50 when semantic evidence is retained: its mean nDCG@10 effect changes from -0.50 to $+0.60$ (Figure 1c; Appendix Table 11). Lexical coverage thus refines semantic predictions rather than serving as an independent substitute.

Corpus evidence complements candidate support. Retaining candidate support improves mean nDCG@10 and both recall metrics over removing it (Appendix Table 12). The support prior thus both weights sources and provides the candidate's query-based starting point, which corpus evidence adjusts rather than replaces.

6 RQ2: When does CPR outperform or complement others?

RQ1 establishes how the components contribute on average. We next examine which query evidence accompanies task-level differences, and whether query expansion changes the benefits of corpus-conditioned prediction. Detailed results appear in Appendix F.

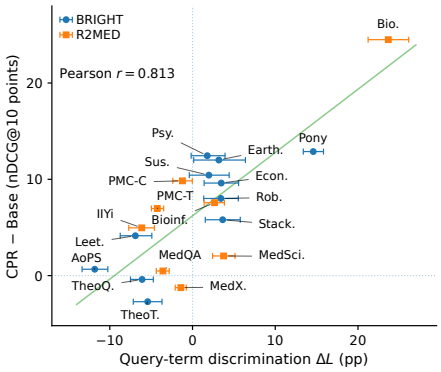


Figure 2: Query-term discrimination and CPR's gain over Base. Horizontal bars show 95% intervals for ΔL; the line is a descriptive linear fit.

Table 2: Recovery and composition. Alone and +CPR report nDCG@10; Rescue and Comb.-only are percentages of Base misses.

Standalone recovery

Method	Rescue	Miss nDCG	ΔHit
HyDE	24.04	8.90	+3.58
QuDAR	15.18	3.81	+2.12
QuDAR-LLM	17.56	4.96	+2.42
CPR	26.47	9.64	+3.01

Composition with expansion

Expansion	Alone	+CPR	Comb.-only
HyDE	26.57	**32.84**	7.20
ThinkQE	21.84	**27.39**	6.89

6.1 Task Evidence and Relative Performance

Query terms may distinguish relevant documents even when they do not fully describe the information need. We measure this distinction using ΔL: mean query-term coverage on labeled relevant documents minus coverage on E5's top-ten competitors not labeled relevant. The descriptor uses relevance judgments and E5 rankings, without consulting CPR outputs. We compare it with each task's mean nDCG@10 gain over Base across the three backbones (Appendix Table 13).

Discriminative query terms accompany larger retrieval gains. Gains rise with ΔL across tasks (Fig 2; Pearson $r = 0.813$). Its association with source-expansion gains is stronger from top-25 to top-100 ($r = 0.771$) than from top-100 to top-1,000 ($r = 0.634$). Beyond top-100, gains are concentrated in Biology and Pony, while other tasks can lose performance (Appendix Fig 4). Query-term discrimination thus accompanies useful expansion without determining a uniformly preferred source scope. We also examine query-absent terminology in supported sources. Greater relevant-side coverage accompanies larger gains from Prior only to Full (Appendix F.4). These label-based diagnostics describe observed differences, not an inference-time selection rule.

6.2 Recovery and Ranking Quality

To distinguish recovery from improvement on successful queries, we separate Base hits and misses by whether the top ten contains a relevant document. Rescue measures recovery among misses, Miss nDCG averages nDCG@10 over all misses, and ΔHit measures the mean change on hits.

Recovery and ranking gains favor different methods. CPR recovers more Base misses and obtains higher Miss nDCG, while HyDE yields greater gains on Base-hit queries (Table 2). Corpus-conditioned prediction thus helps both initially missed and already successful queries, although recovery breadth and ranking quality remain distinct.

6.3 Composition with Query Expansion

The different recovery patterns motivate combining generated query evidence with corpus-conditioned prediction. We apply CPR to HyDE and ThinkQE, our query-augmentation baselines, and compare each composition with the corresponding expansion and standalone CPR. We also count Base misses recovered only by the composition.

Composition recovers queries missed by both standalone methods. Both compositions improve mean nDCG@10 over expansion alone and standalone CPR, and each recovers queries missed by both methods on their own (Table 2; Appendix Table 16). These composition-only recoveries cannot be obtained merely by choosing between the two standalone top-ten lists. Generation changes the available evidence, while CPR uses it to support sources and evaluate candidates. Their combination can therefore address failures shared by the standalone methods.

7 RQ3: How does the retrieval backbone affect CPR?

CPR yields the largest gain over Base with E5, despite its weaker direct retrieval. We compare query–document alignment with source–candidate relations and test the latter under fixed query support. Details appear in Appendix G.

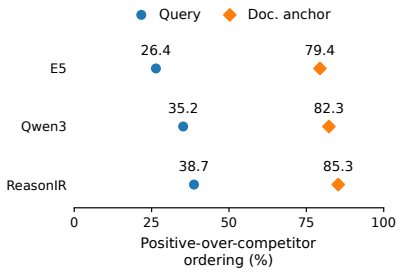


Figure 3: Ordering rates for matched relevant-document–competitor pairs, using the query or an oracle document anchor.

Table 3: Macro nDCG@10 with row evidence fixed. Column backbones supply r_s and u. Bold diagonal entries are Full; Lexical added omits relation prediction.

		Relation source		
Evidence	Lexical	E5	Qwen3	ReasonIR
E5	20.21	**21.40**	22.14	21.99
Qwen3	28.75	30.12	**30.74**	30.73
ReasonIR	26.82	28.71	29.46	**28.96**

7.1 Query Alignment and Document Relations

A backbone may struggle with query alignment even when relevant documents are well connected. We score the same relevant-document–competitor pairs across backbones using the query or another labeled relevant document as an oracle anchor, which is excluded from candidates.

Backbone gaps depend on the connection being evaluated. E5 trails both peers with queries but approaches Qwen3 with document anchors, while ReasonIR leads in both conditions (Figure 3). The E5–Qwen3 query-alignment gap is therefore not reproduced in relations among relevant documents. We next test these relations under query support.

7.2 Corpus Relations under Fixed Query Support

To isolate document relations, we fix one backbone's source and candidate support, lexical evidence, and first result, and replace only $r_s(d \mid z)$ and its corresponding $u(d)$ together using each backbone. Lexical added omits relation prediction, retaining candidate support and lexical evidence.

E5 supplies useful relations despite weaker direct retrieval. Every relation source improves mean nDCG@10 over Lexical added, with Qwen3 highest in each row (Table 3). E5's relations also yield gains under Qwen3 and ReasonIR evidence. The large direct-retrieval gap is not reproduced when only relations change.

Constructing query support contributes to the backbone-specific gains. Support-prior construction accounts for most of each backbone's improvement, with further gains from document prediction. E5 gains more at both stages (Appendix Table 20). Both the available document relations and their use under query support therefore matter.

8 Overall Discussion

Our analyses favor **evaluating retrieval evidence through the predictions it enables, rather than its standalone ranking alone.** A prior both ranks documents and weights sources, while lexical cov-

erage contributes differently when combined with semantic prediction. These findings favor evaluating support construction and relevance prediction together. Evidence scope also differs from output depth, since sources can inform candidates without being returned. In practice, corpus evidence should refine rather than replace the candidate's own query support.

The value of additional evidence varies with the query and collection. **Expanding the available evidence and improving its use are complementary design choices.** Full retains all eligible sources without selecting a source-depth cutoff, while achieving mean nDCG@10 close to top-100. Recovering shared failures through composition supports using generated evidence within corpus-conditioned prediction. Evaluation should distinguish recovery from gains on already successful queries, and recovery rates from ranking quality.

Backbone selection should distinguish direct retrieval from contribution to prediction. Relation exchange shows that a weaker direct retriever can still supply useful corpus relations. Backbones should therefore be assessed by the support priors they construct and the relations they contribute under that support. Better use of corpus evidence thus complements stronger embeddings and generated queries by changing how available information informs document relevance.

9 Related Work

We relate CPR to three approaches for supplementing query evidence in reasoning-intensive retrieval (Su et al., 2025; Wei et al., 2026). Query augmentation uses LLM-generated documents or passages (Gao et al., 2023; Wang et al., 2023a), corpus-guided or iterative expansion (Lei et al., 2024; 2025), and query rewriting or decomposition (Qin et al., 2025; Liu et al., 2025). CPR complements these approaches by applying corpus-conditioned prediction to original or generated evidence.

Corpus-guided retrieval uses retrieved documents for query refinement (Yu et al., 2021; Wang et al., 2023b). ECLIPSE-PRF and RDIME-PRF select dimensions through contrastive feedback and risk estimation (D'Erasmo et al., 2025; 2026). LexBoost combines lexical scores over dense neighborhoods (Kulkarni et al., 2024), while TaxoIndex and SemRank use taxonomy-guided indexing and LLM-selected concepts (Kang et al., 2024; Zhang et al., 2025c). Representation sharpening refines embeddings through LLM-generated contrastive queries (Ashok et al., 2026). Earlier corpus-based methods estimate relevance term distributions (Lavrenko & Croft, 2017), regularize scores through document relations (Diaz, 2005). CPR weights document-relation predictions by query support and compares them with the corpus background.

Evidence aggregation and probabilistic retrieval combine alternative signals or model embedding distributions. QUARK aggregates retrieval scores from generated hypotheses (Lyu et al., 2026), while QuDAR combines original and expanded sparse–dense evidence, with LLM judgments in QuDAR-LLM (Kim et al., 2026). PQR models potential-query embeddings with Gaussian mixtures (Kang et al., 2025), and UEC uses supervised post-hoc Laplace fitting for embedding ensembling (Lim et al., 2026). Parry et al. (2026) combine document priors with query-dependent evidence. CPR differs in representing possible intents with corpus documents and averaging their relevance predictions under query support.

10 Conclusion and Limitations

We introduced CPR (Corpus-conditioned Predictive Retrieval), motivated by the posterior-predictive principle, which uses corpus documents as partial intent representations to refine candidate support through averaged relevance predictions. Across 19 reasoning-intensive retrieval tasks and three retrievers, CPR improves ranking and coverage over direct retrieval without additional training or query generation. Our analyses favor evaluating evidence through the predictions it enables, with corpus evidence complementing query expansion and stronger embeddings. Corpus-wide relations incur storage and computation costs despite reuse across queries. Our source-depth results suggest that restricting prediction to fewer sources could reduce query-time aggregation costs while retaining comparable mean nDCG@10. Future work could extend intent representations from individual documents to complementary document groups and model their joint contribution to relevance prediction. User feedback could further guide support construction across successive searches, allowing corpus evidence to inform retrieval as users refine their information needs.

AI USE STATEMENT

GPT-5.6 Sol was used to assist with equation verification, manuscript writing and editing, and code execution. The research ideas, experimental design, and initial drafts were developed by the authors. All AI-assisted content, code outputs, analyses, and final results were reviewed and verified by the authors.

REFERENCES

Dhananjay Ashok, Suraj Nair, Mutasem Al-Darabsah, Choon Hui Teo, Tarun Agarwal, and Jonathan May. A representation sharpening framework for zero shot dense retrieval. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3735–3751, 2026.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7889–7901, 2023.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Giulio D'Erasmo, Giovanni Trappolini, Fabrizio Silvestri, and Nicola Tonellotto. Eclipse: Contrastive dimension importance estimation with pseudo-irrelevance feedback for dense retrieval. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pp. 147–154, 2025.

Giulio D'Erasmo, Cesare Campagnano, Antonio Mallia, Pierpaolo Brutti, Nicola Tonellotto, and Fabrizio Silvestri. Statistical foundations of dime: Risk estimation for practical index selection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 722–730, 2026.

Fernando Diaz. Regularizing ad hoc retrieval scores. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 672–679, 2005.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1762–1777, 2023.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Junfeng Kang, Rui Li, Qi Liu, Yanjiang Chen, Zheng Zhang, Junzhe Jiang, Heng Yu, and Yu Su. Pqr: Improving dense retrieval via potential query modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13455–13469, 2025.

SeongKu Kang, Yunyi Zhang, Pengcheng Jiang, Dongha Lee, Jiawei Han, and Hwanjo Yu. Taxonomy-guided semantic indexing for academic paper search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7169–7184, 2024.

Joeun Kim, Seunghyouk Yoon, Xuan-Bach Le, Youngeun Nam, Doyoung Kim, Hwanjun Song, and Jae-Gil Lee. Qudar: Query-wise dual-perspective adaptive retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38662–38679, 2026.

Hrishikesh Kulkarni, Nazli Goharian, Ophir Frieder, and Sean MacAvaney. Lexboost: Improving lexical document retrieval with nearest neighbors. In *Proceedings of the ACM Symposium on Document Engineering 2024*, pp. 1–10, 2024.

Victor Lavrenko and W Bruce Croft. Relevance-based language models. In *Acm sigir forum*, volume 51, pp. 260–267. ACM New York, NY, USA, 2017.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations*, volume 2025, pp. 79310–79333, 2025.

Yibin Lei, Yu Cao, Tianyi Zhou, Tao Shen, and Andrew Yates. Corpus-steered query expansion with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 393–401, 2024.

Yibin Lei, Tao Shen, and Andrew Yates. Thinkqe: Query expansion via an evolving thinking process. *arXiv preprint arXiv:2506.09260*, 2025.

Sungjun Lim, Kangjun Noh, Youngjun Choi, Heeyoung Lee, and Kyungwoo Song. Uncertainty-driven embedding convolution. In *International Conference on Learning Representations*, volume 2026, pp. 42164–42199, 2026.

Yaoyang Liu, Junlin Li, Yinjun Wu, and Zhen Chen. Poqd: Performance-oriented query decomposer for multi-vector retrieval. *arXiv preprint arXiv:2505.19189*, 2025.

Rita Qiuran Lyu, Michelle Manqiao Wang, and Lei Shi. Quark: Robust retrieval under non-faithful queries via query-anchored aggregation. *arXiv preprint arXiv:2601.21049*, 2026.

Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

Andrew Parry, Emmanouil Georgios Lionis, Debasis Ganguly, and Sean MacAvaney. Towards a relevance posterior in neural information access. *arXiv preprint arXiv:2607.23561*, 2026.

Xubo Qin, Jun Bai, Jiaqi Li, Zixia Jia, and Zilong Zheng. Reinforced query reasoners for reasoning-intensive retrieval tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21261–21274, 2025.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, et al. Quantifying the reasoning abilities of llms on clinical cases. *Nature Communications*, 16(1):9799, 2025.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, et al. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*, 2025.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Liu Haisu, Quan Shi, Zachary Siegel, Michael Tang, et al. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 48941–48991, 2025.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9414–9423, 2023a.

Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Transactions on the Web*, 17 (1):1–39, 2023b.

Yiyang Wei, Tingyu Song, Siyue Zhang, and Yilun Zhao. A survey of reasoning-intensive retrieval: Progress and challenges. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42101–42122, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

HongChien Yu, Chenyan Xiong, and Jamie Callan. Improving query representations for dense retrieval with pseudo relevance feedback. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 3592–3596, 2021.

Xiangxu Zhang, Lei Li, Xiao Zhou, and Zheng Liu. R2med: A benchmark for reasoning-driven medical retrieval. *arXiv preprint arXiv:2505.14558*, 2025a.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025b.

Yunyi Zhang, Ruozhen Yang, Siqi Jiao, SeongKu Kang, and Jiawei Han. Scientific paper retrieval with llm-guided semantic-based ranking. *arXiv preprint arXiv:2505.21815*, 2025c.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60, 2024.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific data*, 10(1): 909, 2023.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A Method Details

A.1 Semantic Relevance Construction

We start from the normalized reciprocal masses in Equation 9. We retain these masses for the top h non-self neighbors. The neighbor rank law determines h by counting ranks whose normalized mass exceeds a uniform allocation over the $N-1$ non-self documents. When $h < N-1$, we divide the remaining mass equally among the other $N-1-h$ documents.

A.2 Lexical Relevance Estimation

The two lexical factors are computed as

$$\lambda_q(d) = \frac{1}{2}\left(1 + \frac{\sum_{t\in q} w_q(t)\mathbf{1}[t\in d]}{\sum_{t\in q} w_q(t)p_C(t)}\right), \quad \lambda_{\mathrm{src}}(d) = \frac{1}{2}\left(1 + \frac{\sum_{t\in V\setminus q} w_{\mathrm{src}}(t)\mathbf{1}[t\in d]}{\sum_{t\in V\setminus q} w_{\mathrm{src}}(t)p_C(t)}\right) \tag{11}$$

where V is the corpus vocabulary and $p_C(t)$ is the smoothed background term-occurrence rate. We use different weights for the two factors. Query weights $w_q(t)$ emphasize terms occurring in fewer source documents. Source weights $w_{\mathrm{src}}(t)$ emphasize non-query terms whose occurrence increases under $p(z|q)$ relative to uniform source weights.

We estimate the term weights and background occurrence rates from other documents. To do so, we split eligible documents by alternating their ranks under $p(z|q)$ and use the opposite fold for each candidate. Let S denote the source documents in that fold, with source support renormalized within S. We compute

$$p_C(t) = \frac{\sum_{z\in S}\mathbf{1}[t\in z] + 0.5}{|S|+1}.$$

The same fold-specific background is used in the term weights and both coverage factors.

Within the opposite fold S, the query-weighted term-occurrence rate is

$$\mathrm{occ}(t) = \frac{\sum_{z\in S} p(z|q)\mathbf{1}[t\in z]}{\sum_{z\in S} p(z|q)}.$$

We assign query-term weights as $w_q(t) \propto -\log p_C(t)$ for $t \in q$. For non-query terms, we use

$$w_{\mathrm{src}}(t) \propto \begin{cases} \big(\mathrm{occ}(t) - p_C(t)\big) \log \dfrac{\mathrm{occ}(t)}{p_C(t)}, & t \notin q,\ \mathrm{occ}(t) > p_C(t), \\ 0, & \text{otherwise.} \end{cases}$$

Each nonzero weight vector is normalized to sum to one over its term set. Dependence on the candidate's opposite fold is omitted from the notation.

A.3 Document Scoring Derivation

The proportional relevance model in Equation 8 can be instantiated with the common scale $1/N$:

$$p(R_d = 1 | d, z, q) = \frac{r_s(d|z) r_l(d|q)}{N u(d)}.$$

Since

$$r_s(d|z) \le \sum_{z' \in C} r_s(d|z') = N u(d)$$

and $0 \le r_l(d|q) \le 1$, the resulting value lies in $[0, 1]$ and defines a valid binary relevance likelihood. The factor $1/N$ is shared by all candidates and intent sources. Omitting it from the averaged relevance therefore leaves document ordering by Equation 10 unchanged.

B Benchmark Details

BRIGHT. BRIGHT (Su et al., 2025) evaluates retrieval that requires reasoning beyond lexical or semantic matching. It combines StackExchange questions with coding and mathematical tasks based on LeetCode, Pony, the Art of Problem Solving (AoPS) Wiki, and TheoremQA (Chen et al., 2023). Relevant documents include referenced webpages, coding examples, language documentation, solved problems, and theorem statements.

R2MED. R2MED (Zhang et al., 2025a) evaluates Q&A reference retrieval, clinical evidence retrieval, and clinical case retrieval. Biology, Bioinformatics, and Medical Sciences use StackExchange questions. MedXpertQA-Exam, MedQA-Diag, and PMC-Treatment are derived from MedXpertQA (Zuo et al., 2025), MedQA (Jin et al., 2021), and MedRBench (Qiu et al., 2025), respectively. PMC-Clinical uses PMC-Patients (Zhao et al., 2023), while IIYi-Clinical uses clinical cases from the IIYi website.

We use 11 BRIGHT tasks and all eight R2MED tasks, counting their shared Biology task only under R2MED. Table 4 lists the abbreviations and dataset sizes, covering 19 tasks and 2,157 query instances.

C Implementation Details

C.1 Environment

Retrieval and embedding computations use two NVIDIA RTX 6000 Ada Generation GPU with 48 GB of memory. The software environment includes Python 3.11.15, PyTorch 2.11.0+cu130 (CUDA 13.0), NumPy 2.4.4, SciPy 1.17.1, `transformers` 4.51.3, and `sentence-transformers` 5.3.0. BM25 uses Anserini 1.7.1 through Pyserini.

All methods use the same evaluation queries and corpus documents. To compare retrieval procedures on fixed representations, we encode the original queries and corpus once per task–retriever pair and share the cached embeddings across methods. All three retrievers remain frozen. E5-large-v2 uses chunked encoding with the `query:` and `passage:` prefixes, each followed by a space. For query instructions, we follow the settings provided by each benchmark. Within each task–retriever pair, these encoding settings are identical for Base, CPR, and the compared methods. We retain each baseline's retrieval procedure under this shared encoding setup.

Table 4: Task abbreviations and dataset statistics. Document counts refer to the released corpora before query-specific exclusions.

Abbr.	Task	Queries	Documents
BRIGHT			
Earth.	Earth Science	116	121,249
Econ.	Economics	103	50,220
Psy.	Psychology	101	52,835
Rob.	Robotics	101	61,961
Stack.	StackOverflow	117	107,081
Sus.	Sustainable Living	108	60,792
Leet.	LeetCode	142	413,932
Pony	Pony	112	7,894
AoPS	Art of Problem Solving	111	188,002
TheoQ.	TheoremQA-Questions	194	188,002
TheoT.	TheoremQA-Theorems	76	23,839
R2MED			
Bio.	Biology	103	57,359
Bioinf.	Bioinformatics	77	47,473
MedSci.	Medical Sciences	88	34,810
MedX.	MedXpertQA-Exam	97	61,379
MedQA	MedQA-Diag	118	56,250
PMC-T	PMC-Treatment	150	28,954
PMC-C	PMC-Clinical	114	60,406
IIYi	IIYi-Clinical	129	10,449

C.2 LLM-Assisted Baselines

To compare LLM-assisted retrieval procedures without changing the generation model, all LLM components use Qwen3-8B (`qwen/qwen3-8b`) through OpenRouter with thinking disabled. We retain method-specific prompts and the generation settings reported below. Within each method, generated evidence is cached and reused across retrievers, so backbone comparisons use the same generated text.

HyDE. We use the official code. Each query receives eight hypothetical documents from eight independent calls with `n=1`, temperature 0.7, `top_p=1.0`. Frequency and presence penalties are zero, and the stop sequence is three consecutive newline characters. Following the implementation, we average the embeddings of the original query and all eight generated documents.

ThinkQE. We evaluate ThinkQE's iterative retrieval-feedback procedure using the shared Qwen3-8B generator. We use the official implementation with three rounds, `keep_passage_num=5`, `gen_num=2`, `max_demo_len=128`, `query_repeat_weight=3`, and `accumulate=false`. Generation uses temperature 0.7 and `max_tokens=32768`.

QuDAR and QuDAR-LLM. Both variants use Qwen3-8B for passage expansion; QuDAR-LLM also uses it for relevance judgments. These calls use temperature 0.0.

SemRank. We use the official document-level topic and keyphrase extraction procedure. Parsing failures are handled by its retry loop with a new seed.

C.3 LLM-Free Baselines

BM25 and RP. BM25 scores the corpus using the original query text, `DefaultEnglishAnalyzer`, $k_1 = 0.9$, and $b = 0.4$. RP combines dense and BM25 ranks at depth 1,000, assigning rank 1,001 to missing entries.

LexBoost. We retain the official CG transform and `pt.model.add_ranks`, with $k = 16$, $\lambda = 0.7$, and 1,000 candidates.

ECLIPSE-PRF and RDIME-PRF. ECLIPSE-PRF uses two positive and four negative feedback documents, an initial retrieval depth of 1,000, $\alpha = \beta = 1.0$, and retains 50% of embedding dimensions. RDIME-PRF uses $M = 2$ and the estimated-risk threshold.

Document-prior baseline. For Parry et al. (2026), we use `pyterrier-quality/qt5-small` with batch size 100 and the prompt `Document: {} Relevant:`. The candidate depth is 1,000.

C.4 CPR and Composition

CPR scores all eligible documents. The neighbor graph, neighborhood density, and document-specific background $u(d)$ are computed once per task–retriever pair and reused across queries. Each query applies the sparse relation kernel once and performs one lexical pass.

For CPR+HyDE, we reuse the generated documents from standalone HyDE to examine how CPR uses fixed generated evidence. The semantic input is the same mean embedding of the original query and eight hypothetical documents. The lexical input is the union of terms in the original query and those documents. The composition therefore adds no generation calls, while using the existing evidence in CPR's semantic and lexical components.

D Detailed Results

Tables 5 and 6 provide the coverage comparisons at depths 50 and 100. Baseline rows retain the supplied scores.

E RQ1 Experimental Details

All comparisons use 19 tasks (11 from BRIGHT and eight from R2MED) and three retrievers. Biology is counted once, under R2MED. Scores average queries within each task–backbone cell and then weight tasks and backbones equally. We report nDCG@10 and recall on the 0–100 scale. The 95% intervals use 10,000 paired within-task query resamples, keeping each query's three backbone results together.

E.1 Support Prior

We compare Semantic (π_s), Semantic–Lexical (π_{sl}), Separate pool ($(\pi_s + \pi_l)/2$), and CPR pool ($(\pi_s + \pi_{sl})/2$). The lexical-only prior π_l normalizes reciprocal lexical midranks over eligible documents, giving tied scores equal mass. Each condition recomputes semantic predictions, lexical evidence, rank-parity folds, and its retained first document from its own prior. Prior rank orders the support prior directly; Final rank applies the corresponding relevance prediction and document scoring.

E.2 Relevance Likelihood

The source-scope comparison keeps the full eligible output set, candidate support $p(z = d \mid q)$, semantic relations $r_s(d \mid z)$, background $u(d)$, lexical fold assignments, and first result fixed. Source mass is restricted to Base's raw-dense top-k documents and renormalized to recompute semantic predictions and source-supported terminology. The depth sweep uses $k \in \{1, 2, 5, 10, 25, 50, 100, 250, 500, 1000, 2500, 5000, 10000\}$ and Full, which uses all eligible sources. Table 8 reports ranking and coverage comparisons at selected depths.

Source contributions. We partition sources into documents inside and outside the initial dense pool and normalize each group's source mass to one. Each group's semantic prediction is divided

Table 5: Recall@50 (%) on 19 tasks; Avg. is their unweighted mean. Bold/underline mark best/second-best separately in the LLM-free and LLM-assisted (†) groups of each backbone, shown above/below the divider. RP: dense–BM25 rank product.

	BRIGHT											R2MED								
Method	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.	Bio.	Bioinf.	MedSci.	MedX.	MedQA	PMC-T	PMC-C	IIYi	**Avg.**
Lexical																				
BM25	50.99	33.16	30.37	27.80	34.50	35.26	43.38	9.80	15.61	11.57	7.89	36.28	52.34	47.50	2.93	9.39	34.09	48.10	28.36	29.44
E5-large-v2																				
Base	54.04	22.15	22.75	25.95	51.09	25.64	39.04	9.18	16.23	18.24	10.42	22.47	53.13	57.80	10.25	12.17	24.77	38.23	18.91	28.02
RP	63.03	40.34	39.25	34.64	52.31	41.49	48.16	14.41	19.52	21.28	13.05	39.69	65.94	65.83	8.14	13.70	38.41	52.56	27.61	36.81
LexBoost	55.13	35.53	32.82	29.19	34.15	39.06	39.27	11.30	11.81	10.09	10.75	33.18	54.13	55.56	3.45	11.38	36.41	50.95	29.47	30.72
ECLIPSE-PRF	54.89	25.81	22.35	22.86	48.52	29.71	39.78	14.27	7.56	16.28	7.89	27.33	47.53	59.67	10.12	9.57	15.74	31.80	22.98	27.09
RDIME-PRF	54.29	20.21	19.18	23.86	50.59	21.96	37.44	10.12	12.60	15.11	8.99	23.45	47.82	59.07	10.84	13.06	19.72	37.79	19.47	26.61
Parry et al. (2026)	58.30	28.46	25.99	26.49	49.47	33.22	35.09	12.83	16.42	19.63	8.66	33.67	55.31	59.72	**12.36**	14.88	25.88	32.53	15.09	29.68
CPR	**70.53**	**49.61**	**56.25**	**46.87**	**56.48**	**58.87**	**48.84**	**38.47**	**21.91**	**27.59**	**21.60**	**69.72**	**75.44**	**70.83**	11.90	**17.22**	**39.54**	**57.75**	**29.77**	**45.75**
HyDE†	64.95	33.98	40.18	32.19	55.00	39.62	31.94	19.04	17.75	24.90	20.83	54.75	64.11	74.28	34.61	32.94	53.67	41.59	29.69	40.32
SemRank†	38.56	31.11	29.86	27.96	42.41	28.68	**45.59**	9.55	15.78	19.45	17.21	12.88	57.84	61.15	14.15	13.54	28.61	41.67	31.41	29.86
ThinkQE†	65.55	34.59	43.38	36.95	55.78	51.69	34.71	19.92	19.13	26.94	19.74	66.74	72.01	74.56	15.96	26.52	33.30	46.20	35.99	41.03
QuDAR†	64.35	37.60	37.30	34.92	54.09	46.11	37.35	16.57	19.24	26.66	13.05	53.61	72.48	75.33	16.58	22.82	33.34	46.05	31.07	38.87
QuDAR-LLM†	66.23	35.70	40.34	34.43	55.68	48.03	37.11	15.48	17.90	26.77	19.74	60.14	71.77	76.40	17.12	23.47	34.04	44.96	26.19	39.55
CPR+HyDE†	**74.59**	**46.25**	**63.95**	**53.58**	**61.88**	**61.29**	44.17	**54.10**	**28.29**	**38.60**	**40.90**	**86.25**	**82.48**	**83.02**	**36.96**	**38.41**	**63.08**	**61.70**	**50.04**	**56.29**
Qwen3-Embedding-4B																				
Base	54.46	34.66	48.22	30.42	50.30	42.18	59.65	4.13	19.32	**50.68**	57.66	39.31	82.09	75.41	38.36	43.30	67.49	63.60	62.60	48.62
RP	63.85	44.26	49.93	38.97	53.87	47.71	62.35	10.70	22.45	44.00	49.69	47.20	73.74	73.35	31.46	34.67	67.82	69.74	56.95	49.62
LexBoost	53.50	35.66	29.40	29.34	34.81	37.66	40.62	9.60	12.73	9.97	12.06	32.94	56.14	53.61	5.51	11.89	37.39	57.02	32.38	31.17
ECLIPSE-PRF	56.39	35.76	48.35	33.11	52.26	44.34	60.90	2.57	19.34	45.15	59.10	41.65	83.10	77.69	37.26	42.76	67.76	65.20	72.20	49.73
RDIME-PRF	56.57	32.36	45.04	28.93	50.51	38.44	53.58	1.72	18.35	41.68	**60.12**	41.08	82.12	75.44	**39.43**	43.85	66.14	67.25	**72.38**	48.16
Parry et al. (2026)	56.76	39.87	53.32	30.51	49.12	47.98	62.97	4.14	19.96	45.21	52.21	48.12	75.64	72.88	39.31	**44.01**	**72.80**	62.72	65.97	49.66
CPR	**73.94**	**52.35**	**63.33**	**53.83**	**69.05**	**64.14**	**72.93**	**40.51**	**22.60**	47.69	53.74	**70.56**	**86.47**	**79.61**	36.05	40.07	70.03	**73.61**	66.96	**59.87**
HyDE†	54.61	39.54	57.98	31.83	55.20	46.01	60.92	5.75	22.02	53.90	58.69	49.71	83.84	78.82	**45.33**	**48.02**	70.54	70.47	**73.64**	52.99
SemRank†	53.39	37.06	44.10	30.31	46.69	40.12	61.58	4.87	18.80	47.82	56.02	40.77	80.96	74.64	37.61	43.08	62.87	63.74	66.74	47.96
ThinkQE†	49.73	27.49	35.00	25.11	45.58	23.64	65.99	2.72	18.68	49.25	52.65	28.33	77.37	70.53	32.42	35.72	58.21	64.04	64.56	43.53
QuDAR†	57.51	36.62	54.21	30.09	52.14	47.63	54.84	5.67	23.39	49.26	57.36	48.50	83.68	81.07	38.24	44.24	67.59	**70.61**	72.60	51.33
QuDAR-LLM†	58.33	39.03	53.21	31.11	53.87	49.07	54.66	5.02	23.69	49.00	57.47	50.75	83.66	80.97	36.82	43.97	62.40	69.74	71.98	51.30
CPR+HyDE†	**79.06**	**53.47**	**70.46**	**60.70**	**70.47**	**68.49**	**70.82**	**51.51**	**26.72**	**55.53**	**62.31**	**82.94**	**88.93**	**84.97**	39.80	46.55	**73.92**	69.96	66.89	**64.39**
ReasonIR-8B																				
Base	56.01	47.83	59.98	39.34	57.86	53.50	**62.56**	17.78	29.82	**44.38**	51.77	48.30	79.97	74.80	23.92	34.92	56.66	53.51	54.12	49.84
RP	62.70	49.07	56.21	41.90	57.26	50.21	62.01	18.15	30.51	39.65	41.02	57.43	73.96	78.41	22.22	32.10	56.62	**70.25**	51.28	50.05
LexBoost	53.06	35.91	30.58	29.50	35.78	37.48	39.45	11.11	10.88	8.68	13.38	33.11	56.90	55.08	5.51	11.89	36.67	54.61	30.40	31.05
ECLIPSE-PRF	53.39	42.96	54.30	38.90	54.47	50.31	58.80	14.33	30.20	40.86	54.70	52.94	82.59	78.79	29.47	36.38	**62.60**	60.67	61.18	50.41
RDIME-PRF	52.97	44.76	52.26	36.43	56.73	51.44	57.57	15.07	27.24	39.54	**60.95**	51.10	81.79	78.15	32.90	37.00	60.46	65.42	**62.07**	50.73
Parry et al. (2026)	55.86	45.54	65.36	40.00	53.58	56.92	50.48	26.77	29.82	42.24	48.48	66.94	74.10	**79.96**	**33.91**	**41.01**	61.48	60.23	52.95	51.87
CPR	**73.97**	**54.63**	**65.58**	**55.66**	**63.44**	**63.76**	59.07	**35.91**	**31.99**	39.36	46.79	**76.28**	**84.62**	76.62	23.31	33.73	59.27	69.44	55.65	**56.27**
HyDE†	57.20	48.82	67.66	49.22	64.53	57.33	64.03	23.77	26.35	**58.78**	**65.63**	70.55	85.75	**85.83**	**41.93**	**43.46**	**72.38**	66.30	**71.12**	58.98
SemRank†	48.57	42.18	56.60	35.69	54.21	43.81	57.96	24.32	29.39	44.63	47.93	48.53	76.85	76.03	24.04	35.16	51.08	57.31	55.54	47.89
ThinkQE†	56.82	44.00	50.50	33.36	54.58	46.09	64.38	16.11	29.11	57.91	52.32	51.57	79.75	77.39	30.46	32.56	62.06	68.06	55.27	50.65
QuDAR†	57.15	46.98	64.21	43.22	59.45	57.45	**65.87**	23.78	30.11	50.10	60.54	70.20	**86.22**	85.01	31.51	41.77	59.47	66.08	64.79	55.99
QuDAR-LLM†	56.91	45.93	64.05	41.89	60.43	55.62	65.16	23.20	28.57	50.28	60.87	71.87	85.98	84.92	32.06	42.18	58.61	66.08	64.41	55.74
CPR+HyDE†	**75.90**	**51.67**	**72.17**	**60.30**	**69.40**	**66.40**	59.72	**55.48**	**30.14**	50.72	57.66	**86.96**	86.08	83.62	35.83	41.46	69.89	**71.05**	61.61	**62.43**

by $u(d)$. We then measure its average support for qrel-positive documents and for Base's top-ten competitors not labeled relevant (Table 9).

E.3 Document Scoring and Selection

All scoring comparisons fix Full's source prior, semantic and lexical evidence, and first result. Prior only ranks $p(z = d \mid q)$. Semantic added multiplies this support by the query-weighted semantic prediction relative to $u(d)$; Lexical added multiplies it by $r_l(d \mid q)$. Full uses both factors. Retained factors are unchanged across conditions.

Candidate support. An additional control removes only the candidate multiplier $p(z = d \mid q)$, retaining the source weights in the semantic average, lexical evidence, and the first result. Table 12 reports paired differences from removing the candidate multiplier.

F RQ2 Experimental Details

F.1 Task Evidence and Relative Performance

For each query, we measure the fraction of query terms present in each document. The query's discrimination score is the mean coverage on qrel-positive documents minus the mean coverage on E5's raw-dense top-ten competitors not labeled relevant. Averaging across queries gives the task-

Table 6: Recall@100 (%) on 19 tasks; Avg. is their unweighted mean. Bold/underline mark best/second-best separately in the LLM-free and LLM-assisted (†) groups of each backbone, shown above/below the divider. RP: dense–BM25 rank product.

	BRIGHT											R2MED								
Method	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.	Bio.	Bioinf.	MedSci.	MedX.	MedQA	PMC-T	PMC-C	IIYi	**Avg.**
Lexical																				
BM25	59.55	40.77	37.69	37.11	40.93	42.00	50.76	17.22	19.65	15.87	13.38	42.15	61.09	60.02	5.93	11.92	41.23	58.85	37.87	36.53
E5-large-v2																				
Base	60.43	31.94	29.88	28.18	57.94	33.14	47.31	15.38	20.52	20.86	10.42	28.98	65.00	65.64	14.32	19.60	30.74	48.83	24.74	34.41
RP	68.65	47.34	44.93	42.54	57.63	48.99	57.24	24.30	27.83	22.56	15.68	48.92	74.15	76.54	15.04	18.21	44.27	62.35	38.29	43.97
LexBoost	59.86	44.91	39.04	36.93	41.51	46.20	48.39	19.19	15.82	11.47	15.35	40.16	66.25	66.10	4.88	12.35	**45.91**	63.96	37.13	37.65
ECLIPSE-PRF	62.34	33.24	29.44	29.67	59.05	39.61	46.69	23.91	12.65	18.43	11.18	33.62	56.81	69.86	12.60	12.33	19.82	41.96	31.99	33.96
RDIME-PRF	61.72	29.74	27.19	28.27	58.81	30.00	44.67	18.75	19.65	19.10	11.40	26.85	60.90	68.56	14.79	16.48	27.68	46.78	24.41	33.46
Parry et al. (2026)	66.92	39.02	33.83	35.92	57.05	41.88	46.03	21.57	21.16	21.55	12.39	37.31	64.63	69.53	17.16	19.93	33.89	42.98	20.81	37.03
CPR	**77.09**	**59.49**	**66.67**	**60.17**	**67.61**	**70.27**	**59.92**	**57.46**	**30.30**	**34.77**	**29.82**	**79.98**	**84.02**	**82.89**	**19.91**	**24.77**	45.36	**66.37**	**42.58**	**55.76**
HyDE†	72.01	45.40	46.22	34.82	63.54	51.04	42.27	35.44	23.09	27.04	23.36	61.00	76.66	82.91	41.63	41.64	61.03	44.96	43.50	48.29
SemRank†	45.18	38.52	37.84	33.83	49.16	37.85	54.14	16.76	21.63	24.79	25.88	15.30	69.50	69.47	20.99	21.40	34.20	53.07	39.32	37.31
ThinkQE†	72.49	49.20	52.59	46.60	64.41	59.29	43.12	33.72	26.23	31.92	22.92	76.12	77.80	83.41	20.22	31.85	43.14	57.46	48.84	49.54
QuDAR†	72.12	47.82	47.12	41.51	63.88	55.91	48.13	27.32	26.74	31.19	21.05	65.52	78.83	81.81	19.10	29.72	41.06	53.95	43.64	47.18
QuDAR-LLM†	73.25	47.96	47.96	44.95	64.35	57.80	47.43	26.51	26.06	31.96	22.81	71.44	78.93	84.19	19.92	30.03	41.43	53.73	39.50	47.91
CPR+HyDE†	**82.18**	**56.47**	**73.31**	**60.05**	**73.07**	**69.19**	**57.22**	**70.35**	**38.14**	**44.28**	**45.91**	**93.10**	**86.84**	**88.62**	**43.67**	**45.27**	**67.42**	**71.86**	**62.45**	**64.70**
Qwen3-Embedding-4B																				
Base	63.30	49.24	56.56	35.41	60.37	51.92	69.85	9.54	28.70	**56.97**	**62.53**	52.32	87.99	83.24	47.98	**53.72**	76.29	72.22	77.71	57.68
RP	69.24	56.39	58.43	44.09	63.01	55.50	71.47	20.46	31.30	51.61	57.25	58.16	84.77	84.16	39.60	43.50	74.69	77.49	73.37	58.66
LexBoost	59.67	43.88	40.93	36.17	42.28	47.14	45.66	17.28	16.37	12.16	15.35	39.77	68.12	66.16	5.99	12.18	44.68	66.23	41.65	37.98
ECLIPSE-PRF	64.05	49.30	58.69	40.95	61.95	55.44	73.45	6.36	26.80	51.59	60.96	55.04	85.92	82.69	49.27	50.15	76.94	75.95	**83.02**	58.34
RDIME-PRF	63.96	43.25	53.53	34.84	59.74	49.89	63.43	3.85	26.20	48.12	61.54	51.99	87.53	80.38	**50.67**	52.25	73.79	78.51	82.62	56.11
Parry et al. (2026)	67.02	55.76	65.74	36.61	61.54	61.44	67.93	11.94	29.61	49.80	56.83	61.43	84.68	81.34	50.13	53.35	78.18	73.39	79.37	59.27
CPR	**81.59**	**64.85**	**72.12**	**63.35**	**77.41**	**72.67**	**79.80**	**56.74**	**31.56**	55.76	61.43	**81.12**	**91.95**	**88.06**	45.61	51.32	**80.11**	**83.48**	79.37	**69.38**
HyDE†	65.85	53.99	64.95	41.91	64.20	58.01	66.87	11.96	30.87	60.13	67.03	62.78	89.91	87.15	**53.49**	**56.77**	78.51	78.36	**85.48**	62.01
SemRank†	60.48	49.28	57.33	35.45	61.25	47.79	69.42	11.37	26.49	53.45	61.47	54.86	87.10	84.07	47.15	51.52	70.86	73.25	79.11	56.93
ThinkQE†	56.75	37.60	43.92	32.27	59.42	29.88	73.90	6.61	27.37	54.66	60.81	40.36	84.83	75.86	41.98	44.21	69.82	74.20	78.17	52.24
QuDAR†	65.70	54.81	62.51	38.95	62.10	58.82	64.65	12.13	30.41	55.16	66.10	62.84	90.78	90.39	48.41	54.38	74.40	78.65	83.99	60.80
QuDAR-LLM†	65.71	54.71	63.42	39.66	62.74	59.56	64.47	11.95	31.20	55.29	**67.94**	64.55	90.78	89.54	47.50	54.62	71.32	78.44	83.41	60.89
CPR+HyDE†	**84.54**	**66.79**	**76.33**	**66.12**	**79.61**	**75.79**	**78.86**	**65.96**	**35.64**	**61.71**	66.59	**90.74**	**92.65**	**92.04**	51.03	55.28	**80.88**	**82.16**	79.20	**72.73**
ReasonIR-8B																				
Base	65.46	58.31	69.38	47.78	69.60	64.57	69.13	31.18	37.39	**52.24**	64.24	58.15	86.05	83.50	35.71	44.74	65.78	64.40	67.18	59.72
RP	70.60	59.14	66.44	48.50	66.46	60.97	68.08	31.77	36.27	43.94	51.88	67.36	81.53	84.65	31.76	41.70	66.57	74.78	68.15	58.98
LexBoost	59.31	44.99	38.77	37.35	41.05	47.05	47.51	18.84	16.25	11.30	16.01	39.44	68.01	64.40	7.39	12.31	43.79	64.04	39.48	37.75
ECLIPSE-PRF	65.07	55.22	65.43	49.48	65.38	61.96	66.98	24.40	35.31	48.33	67.42	62.96	88.93	**88.53**	40.87	48.98	**71.13**	69.96	**76.23**	60.66
RDIME-PRF	61.84	55.11	62.87	43.98	66.66	57.65	66.75	24.19	33.28	45.13	**70.68**	59.05	88.37	87.28	40.62	50.68	67.64	73.83	73.37	59.42
Parry et al. (2026)	67.18	60.53	71.63	50.47	65.06	67.91	61.21	44.76	38.79	47.18	54.29	77.24	84.47	88.34	**43.40**	**51.22**	66.73	67.18	68.44	61.90
CPR	**80.78**	**66.75**	**74.23**	**65.49**	**72.67**	**72.67**	**70.52**	**58.27**	**39.28**	44.91	53.31	**85.42**	**90.17**	87.41	36.80	47.08	66.72	**79.75**	73.36	**66.61**
HyDE†	67.11	61.33	74.71	56.10	75.56	70.27	**74.61**	40.02	34.72	62.87	**77.80**	82.02	91.12	92.42	**51.62**	**54.81**	**76.54**	73.98	**83.07**	68.46
SemRank†	59.09	54.62	65.30	45.04	66.43	55.28	62.76	38.30	35.31	50.73	57.69	59.47	82.33	84.60	34.93	45.42	58.67	66.81	70.27	57.53
ThinkQE†	64.71	56.40	63.41	43.56	65.16	59.42	71.21	23.94	35.43	**63.82**	62.73	63.46	85.22	86.17	40.24	44.73	69.10	75.80	69.83	60.23
QuDAR†	65.76	60.26	73.04	53.73	71.87	65.88	73.43	38.92	38.43	57.22	68.41	81.59	90.96	92.52	44.80	52.42	67.57	77.49	77.08	65.86
QuDAR-LLM†	68.15	60.94	73.15	54.08	71.84	67.41	72.96	39.34	37.85	56.45	68.73	83.67	**91.13**	91.43	42.94	52.85	67.14	**78.36**	77.08	66.08
CPR+HyDE†	**81.61**	**62.13**	**79.17**	**68.55**	**77.97**	**73.71**	66.78	**71.63**	**40.50**	57.49	66.95	**93.07**	90.87	**92.77**	46.30	52.37	74.83	**78.36**	72.95	**70.95**

Table 7: Support-prior comparisons in macro nDCG@10. Differences are paired; Full denotes CPR pooling followed by document scoring.

Prior	Prior rank	Final rank	Final − Prior	Full − Final [95% CI]
Semantic	23.06	25.77	+2.71	+1.27 [+0.77, +1.76]
Semantic–Lexical	24.01	24.50	+0.49	+2.53 [+2.03, +3.04]
Separate pool	20.51	22.16	+1.66	+4.87 [+4.35, +5.39]
CPR pool	25.76	27.04	+1.27	—

Table 8: Source-scope differences with paired 95% intervals.

Contrast	nDCG@10	Recall@50	Recall@100
Full − Top-100	-0.03 [-0.20, +0.14]	+0.21 [-0.15, +0.57]	+0.40 [+0.00, +0.81]
Full − Top-1,000	-0.16 [-0.25, -0.07]	-0.22 [-0.42, -0.02]	-0.33 [-0.57, -0.09]

level ΔL. Table 13 reports this fixed descriptor and CPR's mean nDCG@10 gain over Base across the three backbones. Their task-level association has Pearson $r = 0.813$ and $R^2 = 0.660$.

Table 9: Background-relative semantic support after normalizing each source group's mass separately.

Sources	Labeled positives	Top-ten competitors
Inside the initial pool	299.7	389.9
Outside the initial pool	19.5	15.6

Table 10: Document scoring with fixed source evidence.

Condition	nDCG@10	Recall@50	Recall@100
Prior only	25.76	51.80	61.43
Semantic added	26.43	53.30	63.49
Lexical added	25.26	51.00	60.95
Full (CPR)	27.04	53.96	63.92

Table 11: Scoring differences with paired 95% intervals. Semantic and Lexical denote the corresponding added-evidence conditions.

Contrast	nDCG@10	Recall@50	Recall@100
Semantic − Prior	+0.67 [+0.36, +0.96]	+1.50 [+0.95, +2.04]	+2.06 [+1.59, +2.51]
Lexical − Prior	-0.50 [-0.83, -0.18]	-0.81 [-1.36, -0.26]	-0.48 [-1.02, +0.05]
CPR − Semantic	+0.60 [+0.32, +0.88]	+0.66 [+0.20, +1.14]	+0.43 [+0.00, +0.86]
CPR − Lexical	+1.77 [+1.50, +2.04]	+2.97 [+2.47, +3.48]	+2.97 [+2.50, +3.45]
CPR − Prior	+1.27 [+0.91, +1.63]	+2.16 [+1.56, +2.76]	+2.49 [+1.94, +3.03]

Table 12: Retaining versus removing candidate support, with paired 95% intervals.

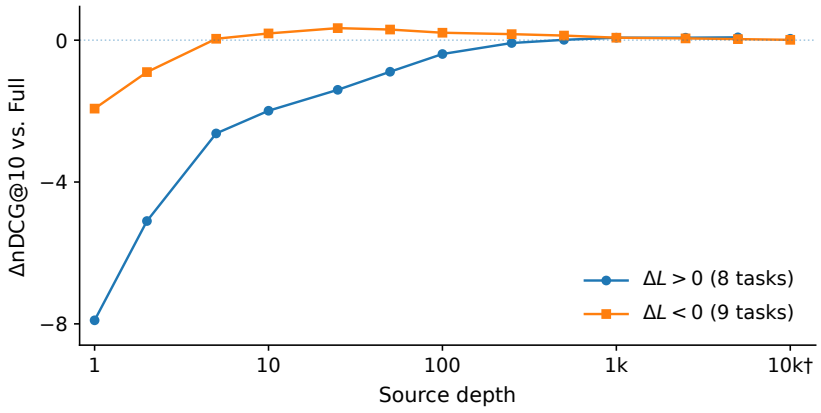
Metric	Full − Without candidate support [95% CI]
nDCG@10	+1.96 [+1.65, +2.27]
Recall@50	+3.65 [+3.06, +4.25]
Recall@100	+4.14 [+3.57, +4.74]

Table 13: Task-level query-term discrimination and retrieval gains. ΔL is in percentage points; CPR − Base is in nDCG@10 points.

Task	ΔL	95% CI	CPR − Base
Earth Science	+3.18	[+0.14, +6.39]	+12.00
Biology (R2MED)	+23.66	[+21.22, +26.13]	+24.50
Psychology	+1.78	[-0.17, +3.93]	+12.45
Pony	+14.59	[+13.39, +15.83]	+12.88
Sustainable Living	+1.97	[-0.39, +4.43]	+10.43
Economics	+3.47	[+1.40, +5.57]	+9.61
StackOverflow	+3.64	[+1.54, +5.77]	+5.80
Robotics	+3.43	[+1.35, +5.53]	+8.01
Medical Sciences	+3.76	[+2.42, +5.14]	+2.04
Bioinformatics	+2.66	[+1.40, +3.85]	+7.57
PMC-Clinical	-1.22	[-2.38, -0.03]	+9.84
MedXpertQA-Exam	-1.41	[-2.09, -0.76]	-1.25
PMC-Treatment	-4.26	[-4.97, -3.49]	+6.97
TheoremQA-Theorems	-5.42	[-7.16, -3.69]	-2.71
MedQA-Diag	-3.58	[-4.38, -2.82]	+0.48
IIYi-Clinical	-6.14	[-7.71, -4.61]	+4.96
TheoremQA-Questions	-6.10	[-7.50, -4.74]	-0.40
LeetCode	-6.91	[-8.74, -4.92]	+4.14
AoPS	-11.82	[-13.35, -10.24]	+0.66

F.2 Source Depth under Different Query Evidence

We group tasks by whether the 95% interval for ΔL lies above or below zero, giving eight positive and nine negative tasks. Psychology and Sustainable Living remain in the overall evaluation but are excluded from both groups because their intervals include zero. We apply the source-scope protocol in Appendix E.2 and average tasks and backbones equally within each group. The association between ΔL and source-expansion gains weakens from top-25 to top-100 ($r = 0.771$), to top-100 to top-1,000 ($r = 0.634$), and to top-1,000 to Full ($r = 0.320$). Full exceeds top-100 by 2.16 points on Biology and 3.26 on Pony, but falls 1.54 points below it on Psychology. Source-depth responses therefore vary across tasks.



† 10k: 7 positive tasks; 9 negative tasks.

Figure 4: Source-depth effects by query-term discrimination. Each curve shows macro nDCG@10 relative to Full within the corresponding group.

Table 14: Selected source-depth comparisons in nDCG@10 with paired 95% intervals.

Group	Source depth	Difference from Full [95% CI]
Positive ΔL	100	-0.39 [-0.71, -0.08]
Positive ΔL	500	+0.01 [-0.20, +0.21]
Negative ΔL	5	+0.04 [-0.29, +0.37]
Negative ΔL	25	+0.34 [+0.09, +0.58]

F.3 Recovery and Composition

A case is a query–backbone pair. A Base hit retrieves at least one labeled relevant document in the top ten; other cases are Base misses. Rescue is the percentage of Base misses recovered by a method. Miss nDCG averages nDCG@10 across all Base misses, including unrecovered cases, and ΔHit is the mean nDCG@10 change on Base hits. These conditional measures average the corresponding cases within each task and then average tasks equally.

Table 15: Recovery and ranking outcomes for the methods compared in the main recovery analysis.

Method	nDCG@10	Rescue (%)	Miss nDCG	ΔHit
HyDE	26.57	24.04	8.90	+3.58
QuDAR	23.35	15.18	3.81	+2.12
QuDAR-LLM	24.04	17.56	4.96	+2.42
CPR	27.04	26.47	9.64	+3.01

Composition. We apply CPR to the generated evidence from HyDE and ThinkQE, the query-augmentation baselines. Comb.-only is the percentage of Base misses recovered by the composition

while both standalone CPR and the corresponding expansion fail to retrieve a relevant document in the top ten. Table 16 reports ranking quality, total rescue, and composition-only recovery.

Table 16: Composition with query expansion. Alone and +CPR report macro nDCG@10; both recovery columns use Base misses as their denominator.

Expansion	Alone	+CPR	Rescue (%)	Comb.-only (%)
HyDE	26.57	32.84	37.83	7.20
ThinkQE	21.84	27.39	29.50	6.89

F.4 Source-Supported Terminology

We examine query-absent terms that distinguish labeled relevant documents from first-pass competitors not labeled relevant. Relevant-side and competitor-side coverage measure the corresponding term weights present in source documents; net coverage is their difference. We compare uniform and support-prior weighting over the same sources, excluding the relevant and competitor documents used to construct the term sets. Prediction gain is the nDCG@10 difference from Prior only to Full.

Across all three retrievers, relevant-side and net coverage have higher observed correlations with prediction gain under support-prior weighting than uniform weighting, while competitor-side coverage is negatively correlated with the gain (Table 17).

We further divide queries into Prior-only score quartiles within each task and compare positive and nonpositive net coverage. For each weighting scheme, we retain only strata containing both groups. Under support-prior weighting, positive net coverage accompanies larger prediction gains for all three retrievers, with the difference intervals above zero. Under uniform weighting, the Qwen3 and ReasonIR intervals include zero (Table 18). These comparisons describe conditional associations rather than causal effects.

G RQ3 Experimental Details

G.1 Query and Document-Anchor Comparisons

The diagnostic uses 1,556 queries with at least two eligible relevant documents; the remaining 601 queries are excluded. Each document anchor is a labeled relevant document held out from the target pairs.

Competitors are selected from the union of the three backbones' raw-dense top-1,000 lists. After removing positives and protocol-ineligible documents, we retain ten by summed reciprocal rank. Anchors, targets, and competitors are shared across backbones. Table 19 reports task-level ordering rates.

G.2 Relation Exchange under Fixed Evidence

The row backbone in Table 3 supplies $p(z \mid q)$ for both source weighting and candidate support, as well as $r_l(d \mid q)$, folds, the eligible set, and the retained first result. Only $r_s(d \mid z)$ and $u(d)$ come from the column backbone; lexical evidence is unchanged.

Lexical added uses $p(z = d \mid q)r_l(d \mid q)$ with the same first result retained. Diagonal entries reproduce Full. Scores average the 19 tasks equally within each row.

Backbone-specific stage changes. Table 20 compares Base, ranking by the support prior, and CPR across all 57 cells. Gains measure Base-to-Prior and Prior-to-CPR changes, with tasks weighted equally.

Table 17: Pearson correlations between source coverage and the nDCG@10 gain from Prior only to Full. Brackets show query-bootstrap 95% intervals.

Source weighting	Relevant-side	Competitor-side	Net
E5			
Uniform	$+0.111$ [0.035, 0.186]	-0.158 [-0.239, -0.075]	$+0.174$ [0.091, 0.252]
Support prior	$+0.316$ [0.249, 0.382]	-0.201 [-0.274, -0.127]	$+0.307$ [0.241, 0.375]
Qwen3			
Uniform	$+0.169$ [0.099, 0.239]	-0.187 [-0.277, -0.095]	$+0.217$ [0.135, 0.295]
Support prior	$+0.287$ [0.216, 0.357]	-0.228 [-0.308, -0.145]	$+0.297$ [0.220, 0.370]
ReasonIR			
Uniform	$+0.112$ [0.048, 0.173]	-0.112 [-0.188, -0.038]	$+0.142$ [0.073, 0.210]
Support prior	$+0.288$ [0.217, 0.356]	-0.222 [-0.297, -0.143]	$+0.299$ [0.230, 0.368]

Table 18: Prediction gains within task-specific Prior-only performance strata. Gains are nDCG@10 differences from Prior only to Full. Only strata containing both net-coverage groups are compared. Brackets show query-bootstrap 95% intervals.

Source weighting	Strata	Net > 0	Net ≤ 0	Difference [95% CI]
E5				
Uniform	14	$+6.22$	$+2.76$	$+3.46$ [1.35, 6.31]
Support prior	14	$+7.24$	-2.72	$+9.95$ [4.74, 16.58]
Qwen3				
Uniform	17	$+3.77$	$+2.74$	$+1.03$ [-1.10, 2.96]
Support prior	18	$+3.89$	-0.14	$+4.04$ [1.52, 7.01]
ReasonIR				
Uniform	16	$+3.47$	$+1.64$	$+1.83$ [-0.10, 4.01]
Support prior	19	$+4.17$	-5.26	$+9.43$ [4.00, 15.78]

Table 19: Positive-over-competitor ordering rates (%). Document anchors are supplied from relevance judgments; the target pairs remain fixed across backbones and input conditions.

	Query			Document anchor		
Task	E5	Qwen3	ReasonIR	E5	Qwen3	ReasonIR
Earth Science	49.4	37.1	39.0	85.5	83.0	86.5
Biology	26.2	29.6	41.0	91.1	89.3	92.0
Psychology	21.5	23.6	41.4	79.5	83.8	91.0
Pony	14.6	13.5	23.2	76.6	72.9	89.0
Sustainable Living	18.3	21.8	35.2	79.9	79.8	87.8
Economics	23.5	22.3	34.1	86.6	90.9	94.7
StackOverflow	42.3	27.3	34.2	86.0	87.4	90.5
Robotics	20.4	16.4	25.6	87.0	83.5	92.4
Medical Sciences	34.4	43.7	45.9	77.1	79.0	81.0
Bioinformatics	29.4	53.5	51.8	82.4	86.3	85.1
PMC-Clinical	33.0	47.3	43.7	66.5	80.9	74.9
MedXpertQA-Exam	12.5	25.4	21.4	88.9	90.6	89.6
PMC-Treatment	23.5	57.8	45.9	82.8	84.7	85.9
TheoremQA-Theorems	31.2	49.4	51.8	83.7	87.9	87.1
MedQA-Diag	14.3	32.3	29.2	91.1	93.5	91.2
IIYi-Clinical	25.2	44.2	41.5	62.4	73.5	69.5
TheoremQA-Questions	21.3	51.9	47.4	78.6	86.8	89.8
LeetCode	40.9	50.5	52.9	55.5	57.5	64.4
AoPS	19.8	21.1	30.4	67.5	73.2	78.5
Macro	26.41	35.19	38.72	79.41	82.34	85.31

Table 20: Stage-wise nDCG@10 and reported changes.

Backbone	Base	Prior	CPR	Prior gain	Prediction gain
E5	12.64	19.51	21.40	+6.87	+1.88
Qwen3	24.10	29.53	30.74	+5.44	+1.21
ReasonIR	24.16	28.24	28.96	+4.08	+0.72