QuDAR: Query-Wise Dual-Perspective Adaptive Retrieval

Joeun Kim[1], **Seunghyouk Yoon**[2], **Xuan-Bach Le**[3],
Youngeun Nam[1], **Doyoung Kim**[1], **Hwanjun Song**[1], **Jae-Gil Lee**[1,*]
[1]Korea Advanced Institute of Science and Technology
[2]Seoul National University [3]Vietnam National University Hanoi
{je.kim, jagil}@kaist.ac.kr

Abstract

Retrieval-augmented generation (RAG) systems depend on retrieval modules to supply grounding evidence for large language models. While hybrid approaches combining sparse and dense retrievers improve performance, most rely on fixed weights that ignore query-specific and corpus-specific variation. Similarly, query expansion has long been used to enrich recall, but its integration with original queries is usually static and can introduce noise. We present QUDAR, a dual-perspective adaptive retrieval framework motivated by a systematic analysis of retrieval behavior across retriever type (sparse vs. dense) and query format (original vs. expanded). Leveraging margin-derived confidence (e.g., top-1–top-2 score gaps) and LLM-based relevance scoring, QUDAR dynamically assigns query-specific weights, enabling effective integration of complementary retrieval signals while mitigating noise. QUDAR is lightweight, retriever-agnostic, and broadly applicable. Experiments show consistent gains over static baselines, improving retrieval quality by 12–16% and yielding more stable performance across queries.

1 Introduction

Retrieval-augmented generation (RAG) have become a widely adopted paradigm for enhancing the factuality and robustness of large language model (LLM) outputs by grounding generation in external document collections (Lewis et al., 2020; Izacard and Grave, 2021). A key determinant of RAG performance is the retrieval component, which selects relevant documents from a large corpus. Numerous studies have shown that improvements in retrieval directly translate to gains in downstream tasks such as question answering and summarization (Guu et al., 2020; Borgeaud et al., 2022; Li et al., 2025). Yet in most systems, retrieval remains static, with fixed configurations that fail to adapt to diverse query formulations or corpus characteristics, limiting overall effectiveness.

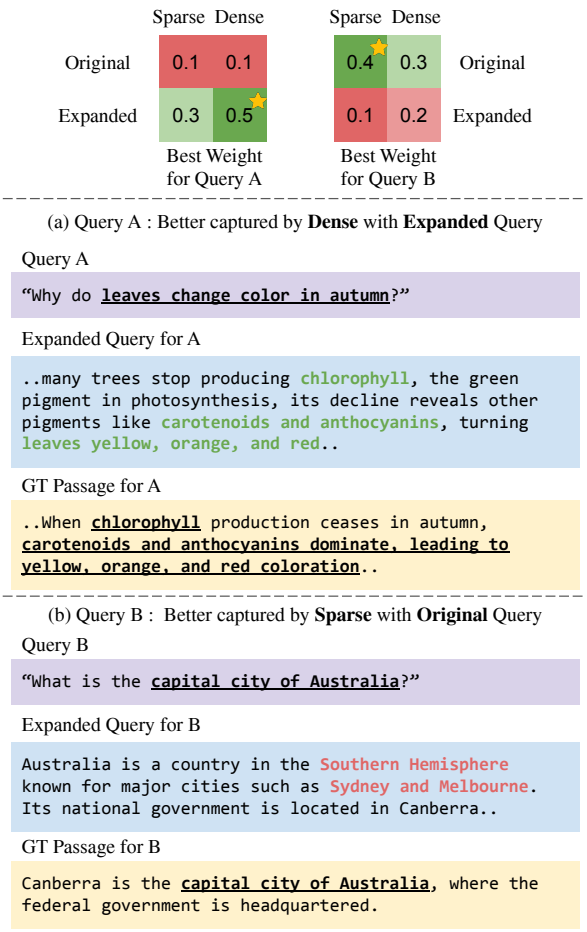

Figure 1: Query-wise dual-perspective adaptation with examples showing different optimal choices across retriever-type and query-format.

This limitation arises because prior RAG frameworks have typically focused on either retriever-type or query-format in isolation. Effective retrieval in RAG requires a *dual-perspective approach* that jointly considers both dimensions. From the retriever perspective, sparse methods such as BM25 (Robertson and Walker, 1994) rank documents by lexical overlap, whereas dense methods encode queries and documents into a shared space and rank documents by semantic similar-

*Corresponding Author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 38662–38679

ity (Karpukhin et al., 2020; Xiong et al., 2021). These methods are complementary, with their relative advantages varying across queries and domains. From the query perspective, information retrieval has long employed query expansion, augmenting the original query with related terms (Carpineto and Romano, 2012). Expanded queries improve recall and semantic coverage for underspecified or domain-mismatched queries, while original queries often better preserve precision by avoiding query drift. Importantly, these two perspectives are not independent. Their interaction determines retrieval behavior, as the effectiveness of a retriever depends on how the query is formulated, and vice versa. In sum, retrieval effectiveness is governed by the interplay between lexical vs. semantic signals and original vs. expanded formulations across these two perspectives.

However, prior work has not fully reflected this dual-perspective interaction. In the retriever dimension, hybrid strategies combine sparse and dense retrieval, using pipelines or score fusion (Santhanam et al., 2021; Cormack et al., 2009; Bruch et al., 2023). While effective, these approaches rely on parameters fixed across queries and corpora, limiting adaptability. In the query dimension, expansion is typically applied by replacing the original query or merging it into a single form (Carpineto and Romano, 2012), where static heuristics collapse their complementary roles. Thus, retriever-type hybridization remains static, and query-type signals are rarely treated as distinct, making it difficult to fully capture and exploit their *combinatorial effect* across both perspectives.

Figure 1 illustrates the interaction between retriever-type and query-format, and their combinatorial effect. *Query A*, "Why do leaves change color in autumn?", benefits from dense retrieval with an expanded query, where domain-specific cues enrich the semantic signal. In contrast, *Query B*, "What is the capital city of Australia?", is better handled by sparse retrieval with the original query, since expansion introduces irrelevant entities and dilutes precision. The optimal weights in Figure 1 highlight this contrast, emphasizing the need for per-query adaptation across both axes rather than static weights.

Therefore, we present a systematic analysis of retrieval hybridization in RAG under a dual-perspective formulation. Our findings show that fixed strategies fail to capture the variability of real queries and corpora. Specifically, we show that (i) retriever-type hybridization cannot be reduced to static pipelines or fixed-weight fusion (§ 3.1); (ii) query-format hybridization, while beneficial, is also constrained when applied statically (§ 3.2); (iii) optimal weights vary widely across queries, indicating substantial headroom for query-wise adaptation (§ 3.3); and (iv) the two perspectives are interdependent, exhibiting a combinatorial effect that static heuristics fail to exploit (§ 3.4).

To instantiate this formulation, we propose **QUDAR** (***Qu***ery-wise ***D***ual-perspective ***A***daptive ***R***etrieval), a framework that enables dual-perspective adaptation through lightweight, *training-free* strategies. We begin with simple rank- and score-fusion baselines (QUDAR-simple), and then introduce adaptive weighting schemes based on margin-derived confidence (QUDAR-confidence) and LLM-based relevance scoring (QUDAR-llm). These methods serve as practical instantiations of the formulation, illustrating how per-query adaptation across both retriever-type and query-format can improve retrieval effectiveness. To the best of our knowledge, this is the first work to jointly consider both perspectives while adapting their contributions on a per-query basis. Our main contributions are as follows:

- We provide a systematic analysis of retrieval hybridization, demonstrating the limitations of static strategies and highlighting the need for per-query adaptation across both perspectives.
- We propose QUDAR, a lightweight and retriever-agnostic framework with three training-free strategies that dynamically weight retrieval signals across both perspectives.
- Experiments across retrieval and RAG benchmarks show 12–16% gains over the best individual retriever and up to 30% improvements over static averaging baselines.

2 Related Work

2.1 Hybrid Retrieval across Retriever Types

To combine the complementary strengths of sparse and dense retrieval, a variety of hybrid strategies have been proposed. The *sequential pipeline* retrieves candidates with a sparse model and reranks them with a dense retriever (Santhanam et al., 2021), while *score- or rank-fusion* combines their outputs using fixed weights or reciprocal rank rules (Cormack et al., 2009; Bruch et al., 2023). While effective, these strategies generally rely on static weights, limiting adaptability. More recently,

Dynamic Alpha Tuning (DAT) (Hsu and Tzeng, 2025) has been introduced to adjust weights per query using LLM-based relevance scoring. However, such query-adaptive methods remain relatively underexplored, and most hybrid retrieval continues to depend on fixed weighting schemes.

2.2 Query Expansion for Retrieval

Query expansion has long been studied as a means to improve recall in information retrieval (Carpineto and Romano, 2012). Classical approaches enrich queries with synonyms or related terms drawn from lexical resources or feedback, while recent methods leverage pretrained language models to generate paraphrases or pseudo-documents. Techniques such as HyDE (Gao et al., 2023) and Query2Doc (Wang et al., 2023b) exemplify this direction, and LLM-based expansion has shown strong performance, particularly in open-domain question answering. However, query expansion is not universally beneficial: additional terms may introduce noise or irrelevant entities, causing *query drift* and weakening retrieval precision. Thus, original and expanded queries play complementary roles, the former better preserving the original query intent, while the latter enriching contextual coverage. Nevertheless, most existing approaches merge the two signals into a single representation or apply them uniformly, with little attention to hybrid strategies and virtually no exploration of query-specific adaptive balancing.

Moreover, prior work on hybrid retrieval has been largely confined to the retriever-type perspective, with little attention to the combinatorial effect of jointly considering query-format perspective. In particular, no existing approaches have explored *query-specific adaptive weighting across both perspectives*. This highlights the need for a unified view that integrates the two perspectives for effective RAG retrieval.

3 Beyond Static Retrieval: Why Hybridization Matters in RAG

We analyze hybrid retrieval along dual perspectives—retriever type and query format—first independently (§ 3.1–§ 3.2), and then jointly at the query level (§ 3.3–§ 3.4).

3.1 Limits of Static Sparse-Dense Balancing

Research Question Existing hybrid retrieval approaches commonly assume a *globally optimal balance* between sparse and dense retrievers. However, such balance may not generalize across datasets or retriever configurations. This naturally raises the question: *Is the optimal sparse-dense weighting consistent across different contexts?*

Experimental Setting We conduct systematic experiments using BM25 as the sparse retriever, paired with two dense retrievers: Contriever (Izacard et al., 2021) and BGE-m3 (Chen et al., 2024). Experiments are performed on four BEIR datasets (FiQA, Climate-FEVER, SciDocs, and HotpotQA). For each setup, we vary the sparse–dense weighting α from 0 to 1 in increments of 0.1, and report performance using Recall@10. Complete numerical results are provided in § A.

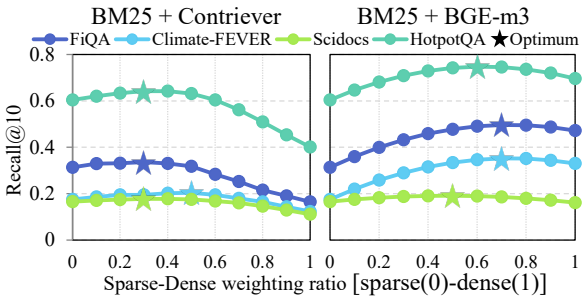

Figure 2: Retrieval performance of sparse-dense hybridization across four datasets using BM25 with Contriever (left) and BGE-m3 (right). The optimal ratio varies across datasets and shifts with retriever choice, indicating sensitivity to both.

Observation Our experiments demonstrate that the optimal sparse–dense weighting ratio is far from universal. Figure 2, the optimal ratio differs markedly across datasets and retrievers. First, there is *dataset sensitivity*: each dataset favors a different balance between sparse and dense signals, reflecting variations in query characteristics and how relevant passages are distributed across the corpus. Second, we find *retriever sensitivity*: when the sparse retriever is fixed, simply switching the dense retriever from Contriever to BGE-m3 consistently shifts the optimal ratio toward more dense-heavy configurations.

Insight These findings suggest that static weighting between sparse and dense retrievers may not hold consistently across contexts. The optimal balance varies with dataset and retriever characteristics, indicating that fixed configurations may not generalize well. This further suggests that hybrid performance depends not only on retriever type but also on its interaction with query characteristics.

3.2 The Double-Edged of Query Expansion

Research Question Query expansion in RAG has typically been applied as substitution or merged into a single query form, rather than treating original and expanded queries as distinct signals. Yet, expansion is inherently *double-edged*: while it can enrich context and improve recall, it may also introduce irrelevant terms and cause query drift. This motivates the following question: *Does hybridizing original and expanded queries actually help?*

Experimental Setting We compare original and expanded queries that generate self-contained passages (Gao et al., 2023), evaluating how their combination affects retrieval performance. Using the same BEIR datasets and retrievers as in § 3.1, we vary the weighting between the two query formats from 0 (original only) to 1 (expanded only) in steps of 0.1 and report Recall@10. Complete numerical results are provided in § A.

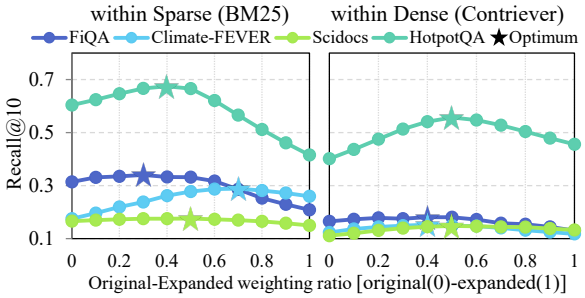

Figure 3: Retrieval performance of query-format hybridization across four datasets using BM25 (left) and Contriever (right). Hybridizing original and expanded queries generally improves retrieval, though the optimal balance varies by dataset and retriever.

Observation Figure 3 (left) shows that expansion effects vary notably across datasets: while Climate-FEVER benefits from using expanded queries, HotpotQA and FiQA perform better with original queries despite identical expansion settings. Across all datasets, however, combining both signals yields stronger performance than relying on either alone. Moreover, the strength of this hybrid effect, as observed in § 3.1, also depends on the specific retriever and dataset used.

Insight Given the double-edged nature of query expansion, effectively leveraging it requires balancing the two signals through hybridization. While combining original and expanded queries generally improves retrieval, the optimal balance varies across datasets and retrievers. This suggests that the effectiveness of query expansion depends on both retriever choice and dataset characteristics.

3.3 Query-Wise Dynamics of Hybrid Weights

Research Question Our earlier analyses showed that the optimal static weighting across retriever type and query format shifts significantly across datasets and retrievers. This raises a natural question: whether such variability also manifests at the query level, and if so, *how much performance gain could be achieved by optimizing hybridization per query beyond static weighting?*

Experimental Setting We estimate query-wise upper bounds (UBs) of hybridization across both retrieval perspectives using the FiQA. For each query, we identify the weighting ratio that maximizes nDCG@10, which better captures ranking quality while showing consistent trends with Recall@10. We then compare the four individual configurations (combinations of sparse vs. dense and original vs. expanded), static hybrids, and dynamic (query-wise optimal) hybrids within each perspective and jointly across both to quantify the potential gains of adaptive hybridization. Results for the remaining datasets are reported in § B.

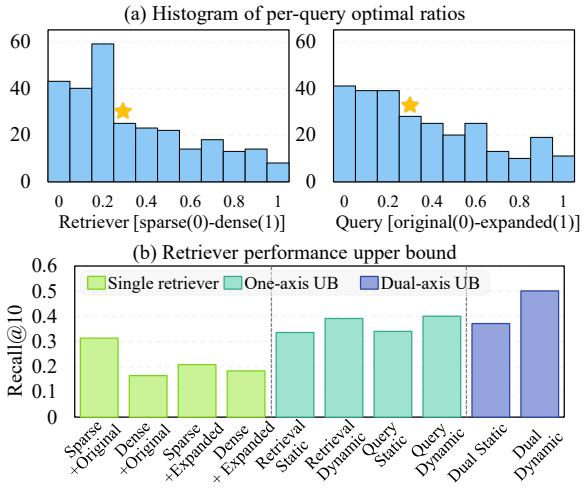

Figure 4: (a) Histograms of per-query optimal weighting ratios for the retriever axis and the query-format axis. (b) Upper-bound retrieval performance across settings: single retrievers, one-axis static/dynamic hybridization, and dual-axis static/dynamic hybridization.

Observation As shown in Figure 4(a), the per-query optimal ratios are widely dispersed, with many queries peaking far from the global static optimum (indicated by a star). This shows that the global optimum, computed as the average ratio across all queries, fails to represent diverse query-specific optima. In Figure 4(b), dynamic hybridization consistently surpasses the static upper bound, achieving approximately 16–18% relative improvement when adapting along a single perspective.

Remarkably, when both perspectives are jointly adapted, the gain increases to nearly 35%, substantially exceeding single-perspective adaptation.

Insight Dual-perspective adaptation yields *synergistic effects*, indicating that robust retrieval benefits from query-wise dynamic weighting across both retriever-type and query-format perspectives. This suggests that hybridization can more fully realize its potential when both perspectives are jointly and adaptively optimized.

3.4 Combinatorial Effects in Dual-Perspective Hybridization

Research Question While dynamic hybridization along a single perspective yields noticeable gains, our analyses show that adapting both perspectives together produces disproportionately larger improvements. This leads to a key question: *Do dual-perspective hybridizations exhibit a combinatorial effect, where their joint adaptation can further unlock the potential of hybrid retrieval?*

Experimental Setting To test whether retriever-type and query-format hybridizations interact combinatorially, we explore a joint optimization space over three parameters: α (sparse–dense), β_1 (original–expanded for the sparse retriever), and β_2 (the same for the dense retriever). Experiments are conducted on the FiQA, with additional results reported in § C. For each query, we exhaustively search for the configuration maximizing nDCG@10. The resulting query-wise optima are aggregated, revealing how the effectiveness of one perspective depends on the other.

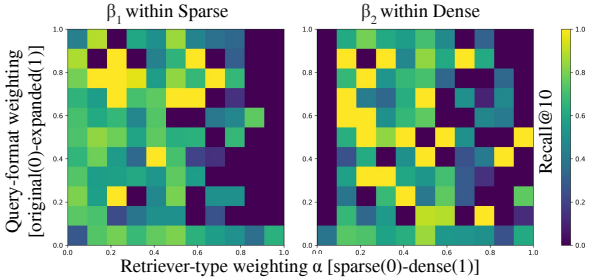

Figure 5: Heatmaps of per-query optimal weights across retriever α and query-format β axes.

Observation Figure 5 shows that retriever-type and query-format hybridization do not operate independently. In some queries, expansion disproportionately strengthens dense retrievers, while in others it favors sparse retrievers, leading to shifts in the optimal retriever balance. Conversely, the relative sparse-dense weighting determines whether expansion helps or harms retrieval performance. These results suggest that the two perspectives are not merely additive but interactively coupled.

Insight The observed interdependence underscores the need to adopt a *dual-perspective* formulation of retrieval. Because the effectiveness of each perspective depends on the other, optimizing either in isolation cannot fully exploit their complementarity. This indicates that hybrid retrieval behavior is inherently *non-additive*, and is better captured when both perspectives are considered jointly, particularly under *query-wise adaptation*.

4 QUDAR : *Qu*ery-Wise *D*ual-Perspective *A*daptive *R*etrieval

4.1 Problem Definition

Given a query q and a document corpus $\mathcal{D}$, we consider dual perspectives that govern retrieval behavior: *retriever type* (sparse vs. dense) and *query form* (original query vs. expanded query). Applying both perspectives yields four retrieval outputs:

$$
\begin{aligned}
R_{OS}(q) &= \text{Retriever}_{\text{Sparse}}(q),\\
R_{OD}(q) &= \text{Retriever}_{\text{Dense}}(q),\\
R_{ES}(q) &= \text{Retriever}_{\text{Sparse}}(QE(q)),\\
R_{ED}(q) &= \text{Retriever}_{\text{Dense}}(QE(q)),
\end{aligned}
$$

where R_{OS} and R_{OD} denote the ranked document lists produced by sparse and dense retrievers respectively when applied to the original query, and R_{ES} and R_{ED} denote their counterparts using the expanded query. $R_i(q)$ provides both a ranking and associated document scores. We denote by $\text{rank}_i(d)$ the position of document d in $R_i(q)$, and by $\text{score}_i(d)$ its normalized retrieval score.

Our goal is to assign query-specific weights

$$
w_{OS}, w_{OD}, w_{ES}, w_{ED} \in [0,1], \qquad \sum_i w_i = 1,
$$

that determine the contribution of each list to the final hybrid retrieval result. Formally, we define the hybrid retrieval function as

$$
R_{\text{hybrid}}(q) = \sum_{i \in \{OS,OD,ES,ED\}} w_i \cdot R_i(q).
$$

The objective is to find the optimal weight

$$
\mathbf{w}^* = \arg\max_{\mathbf{w}} \ \mathcal{S}(R_{\text{hybrid}}(q)), \tag{1}
$$

where $\mathcal{S}(\cdot)$ denotes a retrieval scoring function that estimates the quality of the retrieved set.

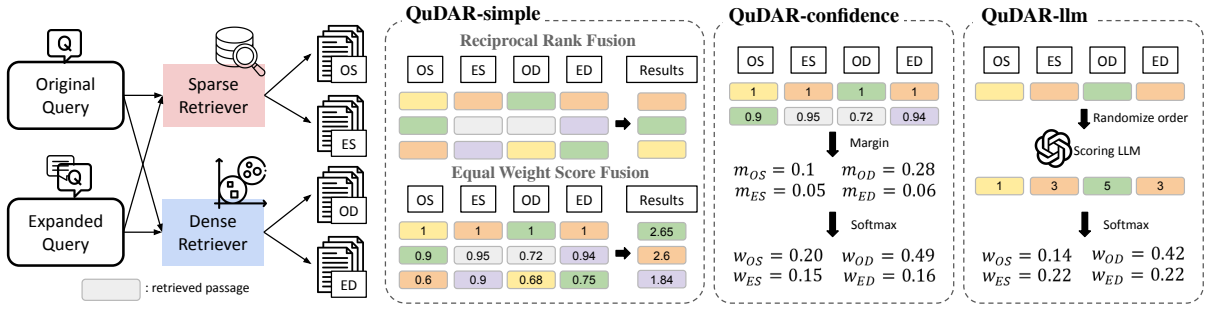

Figure 6: Overall framework of QUDAR. QUDAR integrates four retrieval signals to perform query-wise dual-perspective hybridization. It supports three weighting strategies: **QUDAR-simple**, **QUDAR-confidence** (confidence-based weighting), and **QUDAR-llm** (LLM-based scoring). The passages with the same color denote identical retrieved passages across different signals.

4.2 Overall Framework

Figure 6 illustrates the overall pipeline of QUDAR. Given the four retrieval outputs defined in § 4.1, the framework integrates them through weighted aggregation to produce a final hybrid ranking. For each query, QUDAR applies a weighting mechanism to combine the signals. We consider three variants that differ in how the weights are determined: **QUDAR-simple** uses fixed aggregation rules, **QUDAR-confidence** estimates weights based on retrieval confidence, and **QUDAR-llm** leverages LLM-based relevance scoring. All three share the same dual-perspective formulation, differing only in how they estimate the relative contribution of each signal.

4.3 QUDAR-simple

We begin with the most straightforward form of hybridization. QUDAR-simple combines the four retrieval outputs (OS, OD, ES, ED) without query-specific adaptation or external supervision. Despite its simplicity, this approach already captures both retrieval perspectives and thus serves as a meaningful reference point.

QUDAR-simple (RRF). A rank-based fusion alone can yield hybridization effects. Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) prioritizes documents ranked highly by any retriever, while smoothing contributions across all four lists. As a result, it preserves strong signals from individual retrievers while mitigating noise, allowing complementary results to reinforce each other.

$$\text{score}_{\text{RRF}}(d) = \sum_{i \in \{OS,OD,ES,ED\}} \frac{1}{k + \text{rank}_i(d)},$$

where $\text{rank}_i(d)$ is the position of document d under retrieval setting i, and k is a smoothing parameter.

QUDAR-simple (Equal-weight). While RRF exploits only rank information, equal-weight fusion directly averages normalized retrieval scores. This approach incorporates score distributions, thereby reflecting relative retriever confidence rather than discarding it. By equally weighting all four retrieval outputs, it provides a balanced aggregation across retriever and query-format perspectives.

$$\text{score}_{\text{Equal}}(d) = \frac{1}{4} \sum_{i \in \{OS,OD,ES,ED\}} \text{score}_i(d).$$

4.4 QUDAR-confidence

We introduce QUDAR-confidence, a training-free adaptive weighting method that infers retrieval reliability from the confidence margin of each retrieval signal. The core idea is that a retriever should be trusted more when it ranks its top candidate decisively higher than the rest, indicating higher confidence and lower uncertainty.

For each retrieval list $i \in \{\text{OS}, \text{OD}, \text{ES}, \text{ED}\}$, we define the confidence margin as

$$m_i = \max(\text{score}_{i,1} - \text{score}_{i,2}, 0),$$

where $\text{score}_{i,1}$ and $\text{score}_{i,2}$ denote the normalized scores of the top-1 and top-2 passages, respectively.

A larger margin implies a more confident ranking signal, while a smaller margin suggests ambiguity among top candidates. The margins are then transformed into normalized reliability weights through a temperature-controlled softmax:

$$w_i = \frac{\exp\left(\frac{m_i+\epsilon}{\tau}\right)}{\sum_j \exp\left(\frac{m_j+\epsilon}{\tau}\right)},$$

where τ controls smoothness and ϵ prevents zero-weight collapse. These weights are applied to the four retrieval signals to perform query-wise adaptive fusion.

Dataset			FiQA			SciDocs		
Methods \ Metrics			Recall@10	nDCG@10	Precision@1	Recall@10	nDCG@10	Precision@1
Individual Retriever	Original-Sparse		0.3135	0.2500	0.2299	0.1655	0.1578	0.1810
	Original-Dense		0.1649	0.1299	0.1173	0.1112	0.1031	0.1160
	Expanded-Sparse		0.2085	0.1496	0.1127	0.1497	0.1370	0.1460
	Expanded-Dense		0.1835	0.1363	0.1049	0.1487	0.1390	0.1570
Single-Perspective	Retriever-Type	Static Avg	0.2787	0.2232	0.2044	0.1597	0.1504	0.1703
		Static UB	0.3361	0.2718	0.2500	0.1776	0.1683	0.1970
		DAT	0.2842	0.2332	0.2284	0.1614	0.1525	0.1720
	Query-Format	Static Avg(S)	0.2979	0.2303	0.1939	0.1682	0.1583	0.1759
		Static UB(S)	0.3404	0.2726	0.2423	0.1757	0.1666	0.1880
		Static Avg(D)	0.1647	0.1262	0.1063	0.1362	0.1289	0.1510
		Static UB(D)	0.1810	0.1406	0.1265	0.1479	0.1413	0.1750
Dual-Perspective	Static Avg		0.2779	0.2143	0.1806	0.1715	0.1615	0.1820
	Static UB		**0.3715**	**0.2915**	**0.2562**	0.1917	0.1813	**0.2060**
	QUDAR-simple(RRF)		0.3219	0.2461	0.1944	0.1893	0.1757	0.1950
	QUDAR-simple(Equal)		0.3402	0.2654	0.2284	0.1906	0.1792	0.2000
	QUDAR-confidence		0.3438	0.2684	0.2330	**0.1920**	0.1799	0.1990
	QUDAR-llm		0.3639	0.2837	0.2454	**0.1920**	**0.1816**	0.2050

Table 1: Retrieval results on two datasets: FiQA and SciDocs. The best results are **in bold**, and the second-best results are underlined.

4.5 QUDAR-llm

We further propose QUDAR-llm, a training-free adaptive weighting method that leverages LLM-based relevance scoring to infer query-specific reliability among the four retrieval signals. This approach is inspired by DAT (Hsu and Tzeng, 2025), which adjusts retriever-type weights in a training-free manner; however, QUDAR-llm extends this idea to handle the joint interaction between retriever-type and query-format.

For a given query q, the top-1 passages retrieved from the four retrieved lists are presented to the LLM along with the query text. The LLM evaluates how sufficiently each passage can answer the query and assigns a relevance score from 0 to 5 for each passage. We use `gpt-4o-mini` for scoring, and the full evaluation prompt is provided in § D. To prevent any bias arising from retriever-type or presentation order, the passages are randomly shuffled before the scoring, and their scores are later remapped to the original retrieval sources.

The obtained scores $\text{score}_i^{\text{LLM}}$ are converted into probabilistic weights using a softmax function:

$$w_i = \frac{\exp\left(\frac{\text{score}_i^{\text{LLM}}+\epsilon}{\tau}\right)}{\sum_j \exp\left(\frac{\text{score}_j^{\text{LLM}}+\epsilon}{\tau}\right)},$$

where τ is a temperature parameter controlling the smoothness of the weight distribution.

By directly assessing how well each retrieved passage addresses the query, QUDAR-llm estimates semantic reliability, enabling query-wise dual-perspective hybridization.

5 Evaluation

5.1 Experimental Setup

Datasets We evaluate our method on five datasets: four from the BEIR benchmark (Thakur et al., 2021)-FiQA, SciDocs, HotpotQA, and Climate-FEVER. These datasets cover diverse domains and query types, including multi-hop reasoning and domain-specific retrieval. Further dataset statistics and details are provided in § E.

Baselines We employ BM25 (Robertson and Walker, 1994) as the sparse retriever and Contriever (Izacard et al., 2021) as the dense retriever across all experiments. Expanded queries are generated using `LLaMA 3.1-8B`, following prior work (Gao et al., 2023; Zhang et al., 2024) that reformulates queries into *self-contained passages*-an effective strategy shown to enhance recall and semantic coverage. We evaluate alternative retrieval components and query expansion methods, with detailed results in § F. The full prompt used for query expansion is provided in § G. Our code and implementation details are available at `https://github.com/kaist-dmlab/QuDAR`.

We compare QUDAR against *three categories* of baselines, grouped by hybridization perspective.

1. **Individual retrievers:** Using only one of the four retrieval settings: Original-Sparse (OS), Original-Dense (OD), Expanded-Sparse (ES), and Expanded-Dense (ED).
2. **Single-perspective hybrids:**
   - Retriever-type hybrids (original queries only): Static averaging, grid-searched static weight-

Dataset			Climate-FEVER			HotpotQA		
Methods \ Metrics			Recall@10	nDCG@10	Precision@1	Recall@10	nDCG@10	Precision@1
Individual Retriever	Original-Sparse		0.1750	0.1376	0.1160	0.6039	0.5741	0.6663
	Original-Dense		0.1235	0.0909	0.0606	0.4011	0.3531	0.3749
	Expanded-Sparse		0.2598	0.2053	0.1720	0.4151	0.3605	0.3625
	Expanded-Dense		0.1960	0.1624	0.1564	0.4838	0.4375	0.4675
Single-Perspective	Retriever-Type	Static Avg	0.1784	0.1420	0.1216	0.5730	0.5359	0.6097
		Static UB	0.2040	0.1651	0.1498	0.6419	0.6096	0.7025
		DAT	0.1699	0.1375	0.1251	0.5781	0.5567	0.6641
	Query-Format	Static Avg(S)	0.2506	0.2020	0.1784	0.5869	0.5397	0.5898
		Static UB(S)	0.2886	0.2346	0.2117	0.6746	0.6342	0.7134
		Static Avg(D)	0.1379	0.1080	0.0911	0.4943	0.4439	0.4743
		Static UB(D)	0.1527	0.1195	0.1075	0.5570	0.5013	0.5357
Dual-Perspective	Static Avg		0.2169	0.1761	0.1603	0.6024	0.5529	0.6001
	Static UB		**0.2972**	**0.2408**	**0.2150**	**0.7110**	**0.6694**	0.7463
	QUDAR-simple(RRF)		0.2411	0.1931	0.1661	0.6741	0.6114	0.6460
	QUDAR-simple(Equal)		0.2464	0.2045	0.1967	0.6924	0.6423	0.6948
	QUDAR-confidence		0.2473	0.2050	0.1967	0.6962	0.6478	0.7047
	QUDAR-llm		0.2583	0.2138	0.2085	0.6883	0.6579	**0.7515**

Table 2: Retrieval results on the other two datasets: Climate-FEVER and HotpotQA (same notation as Table 1).

ing, and Dynamic Alpha Tuning (DAT) (Hsu and Tzeng, 2025), a query-adaptive method that dynamically adjusts the balance between sparse and dense retrievers.

- Query-format hybrids (retriever fixed): Static averaging, grid-searched weighting.

3. **Dual-perspective hybrids:** Static averaging, grid-searched weighting, and QUDAR, a *query-wise dual-perspective* hybrid approach encompassing three variants (QUDAR-simple, QUDAR-confidence, and QUDAR-llm).

Metrics We evaluate retrieval performance using three standard metrics: Recall@10, nDCG@10, and Precision@1. Recall@10 measures the proportion of relevant documents retrieved, while nDCG@10 accounts for both relevance and ranking position. Precision@1 reflects top-ranked retrieval accuracy.

5.2 Main Results

5.2.1 Retrieval Performance

Tables 1 and 2 summarize the results for the three baseline categories on the four datasets.

QUDAR outperforms all individual retrievers across datasets. Compared to the strongest individual retriever, it achieves on average 12–16% higher recall, and improves precision on CLIMATE-FEVER by up to 21%. The performance variation across datasets indicates that retrieval effectiveness depends heavily on both the retriever type and the query expansion strategy. This suggests that a query-wise dual-perspective adaptation, as implemented in QUDAR, can complement these variations and yield consistently robust performance across domains.

In the single-perspective setting, QUDAR consistently surpasses both static and dynamic baselines. On the retriever-type, it achieves 28% gains over static averaging, 8% gains over the grid-searched upper bound, and notably exceeds DAT by more than 25% on average. On the query-format, it yields 3–22% improvements over static averaging and up to 9% over the grid-searched upper bound. While single-perspective adaptation can partially capture hybridization benefits, the results vary with dataset characteristics and the effect of query expansion, highlighting the need for per-query adaptive integration across both perspectives.

QUDAR also achieves comparable performance to static dual-perspective hybrids. It improves recall by up to 30% and precision by up to 35% relative to the static hybrid average. Although its performance is slightly below the grid-searched upper bound on some datasets, it maintains over 90% of that performance in most cases and even surpasses it on SCIDOCS. Unlike grid search, which requires exhaustive evaluation of all retrieval combinations and is thus highly time-consuming and impractical, QUDAR dynamically determines query-wise weights without any training or search, achieving both effectiveness and efficiency.

5.2.2 Generation Performance

To evaluate whether improvements in retrieval quality translate into downstream RAG performance, we conduct end-to-end QA experiments on FIQA and HOTPOTQA. Due to the cost of generation-

Group	Method	FiQA		HotpotQA	
		Recall@10	Gen.	Recall@10	Gen.
Individual Retriever	Original-Sparse	0.3133	0.14	0.615	0.36
	Original-Dense	0.1613	0.10	0.411	0.28
	Expanded-Sparse	0.2033	0.12	0.411	0.31
	Expanded-Dense	0.1349	0.10	0.463	0.33
Single-Perspective	Retriever-Type Fusion	0.3329	0.15	0.656	0.40
	Query-Format Fusion	0.1774	0.10	0.557	0.39
Dual-Perspective	QUDAR-simple (RRF)	0.3153	0.16	0.668	0.42
	QUDAR-simple (Equal)	0.3372	0.16	0.685	0.43
	QUDAR-confidence	0.3399	0.17	**0.687**	**0.44**
	QUDAR-llm	**0.3589**	**0.18**	0.683	**0.44**

Table 3: Retrieval and generation performance on FiQA and HotpotQA(sampled subsets). "Gen." denotes generation quality measured by an LLM-based metric. The best results are **in bold**.

Dataset	Time (s/query)	Tokens (#/query)	Cost ($/query)
FiQA	0.69	1271	1.9e-4
SciDocs	0.70	1145	1.7e-4
Climate-Fever	0.75	1263	1.9e-4
HotpotQA	0.79	636	9.5e-5

Table 4: Cost and latency of LLM-based weighting using `gpt-4o-mini`.

based evaluation, we sample 500 queries from each dataset. For each method, retrieved passages are provided to an LLM to generate answers, which are then evaluated using an LLM-based metric (Wang et al., 2023a). We use `Qwen2.5-7B` for both generation and evaluation, with identical settings across methods. Results are summarized in Table 3.

QUDAR consistently improves generation performance. Across both datasets, improved retrieval quality translates into higher generation quality. QUDAR achieves over 20% gains compared to the best individual retriever, and further improves over single-perspective methods by approximately 6% on FIQA and 10% on HOTPOTQA. These results show that jointly modeling retriever type and query format leads to more effective context selection and better answer quality.

5.3 Cost and Latency Analysis

QUDAR-llm requires an additional LLM call per query, introducing computational cost and latency. To quantify this overhead, we measure inference time, token usage, and monetary cost using `gpt-4o-mini` across all datasets (Table 4). Each query requires approximately 0.7 seconds of inference and several hundred to over a thousand tokens, resulting in additional cost. This overhead may be a concern in real-time or large-scale settings. QUDAR also includes training-free variants (QUDAR-simple and QUDAR-confidence) that avoid LLM calls while still outperforming single-perspective baselines. This allows the choice of variant to be adapted based on the desired trade-off between efficiency and performance.

As shown in § 3, the dual-perspective formulation exposes greater performance headroom than single-perspective methods. While LLM-based scoring provides strong relevance signals, the gains of QUDAR-llm arise from combining complementary signals across retriever type and query format. QUDAR-llm explores part of this upper bound, indicating that the dual-perspective formulation enables more effective retrieval combinations. Additional analysis of LLM-based weighting is provided in § H.

6 Conclusion

We present QUDAR (***Qu***ery-wise ***D***ual-perspective ***A***daptive ***R***etrieval), a lightweight and retriever-agnostic framework that adaptively integrates retrieval signals across two complementary perspectives: retriever-type (sparse vs. dense) and query-format (original vs. expanded). Through extensive analysis, we highlighted the limitations of static hybridization and the need for per-query adaptation along both perspectives.

QUDAR realizes this principle through three training-free variants: QUDAR-simple, QUDAR-confidence, and QUDAR-llm. Across multiple retrieval and RAG benchmarks, QUDAR achieves consistent gains, 12–16% over the best individual retriever and up to 30% over static averaging baselines, demonstrating strong generalization across domains and query types. These results suggest that *query-wise dual-perspective adaptation* is an effective strategy for retrieval optimization.

Limitations

Query-wise weight adjustment is expected to depend on the characteristics of the query, corpus, and the chosen retrievers. In this work, we explored several training-free strategies that adjust weights using signals captured from four retrieval combinations. While our analysis demonstrates that dynamic weighting per query has significant potential for performance gains, our methods do not fully reach the upper-bound performance. Incorporating learned predictors or training-based adaptive weighting could further narrow this gap.

Moreover, our results show that retrieval performance is highly sensitive to both retriever type and query formulation. While we explored a range of retrieval components and query expansion methods (§ F), these do not exhaust the space of possible combinations. Further exploration of additional retrieval architectures and expansion strategies could provide a more comprehensive understanding of the generality of these findings.

Acknowledgments

This research was partly supported by the Korea Institute of Science and Technology Information (KISTI) in 2026 (No. (KISTI)K26L3M1C1, 50%), aimed at developing KONI (KISTI Open Neural Intelligence), a large language model specialized in science and technology, and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220157, Robust, Fair, Extensible Data-Centric Continual Learning, 50%).

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2206–2240.

Sebastian Bruch, Siyu Gai, and Amir Ingber. 2023. An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems*, 42(1):1–35.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1–50.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv*, 2402.03216.

Gordon V Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 758–759.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2288–2292.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1762–1777.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval-augmented language model pre-training. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3929–3938.

Hsin-Ling Hsu and Jengnan Tzeng. 2025. Dat: Dynamic alpha tuning for hybrid retrieval in retrieval-augmented generation. *arXiv*, 2503.23013.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv*, 2112.09118.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 874–880.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Siran Li, Linus Stenzel, Carsten Eickhoff, and Seyed Ali Bahrainian. 2025. Enhancing retrieval-augmented generation: A study of best practices. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 6705–6717.

Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, volume 27, pages 232–241.

Joseph John Rocchio Jr. 1971. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*.

Keshav Santhanam, Omar Khattab, Da Zheng, Ji Ma, and Christopher Re. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 817–832.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023a. Evaluating open-qa evaluation. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, pages 77013–77042.

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

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

Lee Xiong, Chenyan Wu, Jingfei Liu, William Yang Wang, Mo Yu, Xiaoxiao Guo, Zhiyuan Gao, Nikhil Mallinar, Saurabh Tiwary, Rangan Majumder, et al. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–14.

Le Zhang, Yihong Wu, Qian Yang, and Jian-Yun Nie. 2024. Exploring the best practices of query expansion with large language models. *arXiv*, 2401.06311.

A Detailed Results of Retriever-Type and Query-Format Hybridization

We provide complete numerical results for all hybridization analyses. Table 5 reports performance when varying the sparse-dense weighting α from 0 to 1 for the CONTRIEVER retriever. Tables 6 and 7 show the results of query-format hybridization—varying the ratio between original and expanded queries—under sparse and dense retrievers, respectively. The experiments reveal that the optimal sparse-dense weighting ratio varies substantially across datasets and retrievers. The relative importance of sparse and dense signals differs by dataset, and even under identical settings, the optimal balance shifts depending on retriever characteristics. Query expansion also exhibits dataset-dependent behavior: it improves performance on CLIMATE-FEVER, whereas HOTPOTQA and FIQA favor original queries. These results indicate that hybrid performance cannot be maintained by any fixed weighting scheme and emphasize the need for adaptive weighting that accounts for both query and retriever characteristics.

B Dual-Perspective Upper Bound

To complement the experiments in § 3.3, we estimate query-wise upper bounds (UBs) of hybridization for the remaining BEIR datasets: Climate-FEVER, SciDocs, and HotpotQA. For each query, we identify the weighting ratio that maximizes nDCG@10, exploring both retrieval perspectives: retriever-type (sparse vs. dense) and query-format (original vs. expanded). We vary the weighting from 0 to 1 in increments of 0.1. The queries showing identical performance across five or more ratios are excluded, because hybridization offers no meaningful effect in such cases. For the queries with multiple optimal ratios, we take their mean value as the final optimum.

Figures 7–9 illustrate the distribution of optimal ratios for both perspectives, revealing substantial variation across datasets. Some lean toward sparse signals, while others favor dense or expanded queries, highlighting dataset- and retriever-specific sensitivity. Table 8 reports the overall performance under static and dynamic UB settings for each perspective. The results show consistent improvements of dynamic weighting over static combinations, quantitatively supporting the need for per-query adaptive hybridization across both retrieval perspectives.

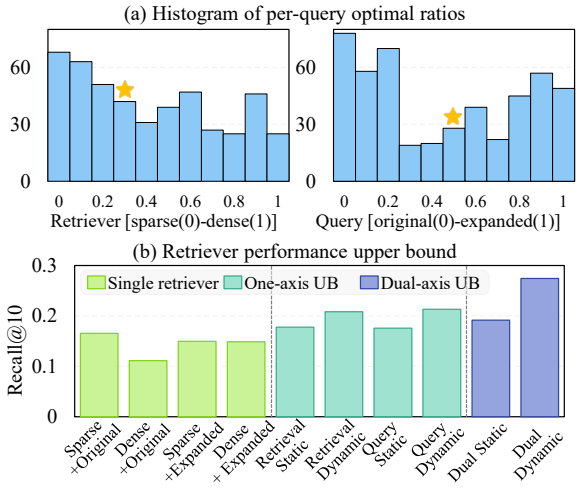

Figure 7: Distributions of per-query optimal weighting ratios (a) and upper-bound retrieval performance (b) in Scidocs.

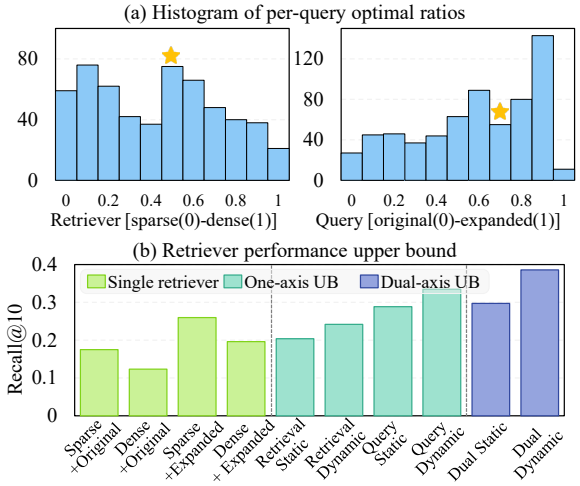

Figure 8: Distributions of per-query optimal weighting ratios (a) and upper-bound retrieval performance (b) in Climate-FEVER.

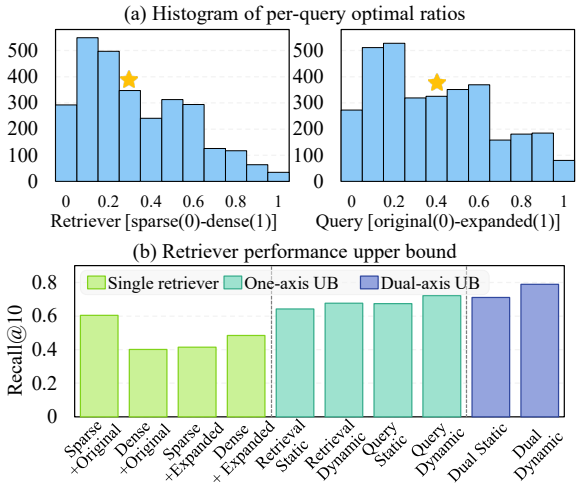

Figure 9: Distributions of per-query optimal weighting ratios (a) and upper-bound retrieval performance (b) in HotpotQA.

Dataset	Metrics \ Ratio	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FiQA	Recall@10	0.3135	0.3293	0.3316	0.3361	0.3304	0.3184	0.2838	0.2522	0.2155	0.1905	0.1649
	nDCG@10	0.2500	0.2617	0.2680	0.2718	0.2634	0.2515	0.2267	0.2020	0.1765	0.1535	0.1299
	MAP@10	0.1895	0.1986	0.2060	0.2087	0.1996	0.1888	0.1693	0.1497	0.1299	0.1111	0.0916
	Precision@1	0.2299	0.2392	0.2469	0.2500	0.2361	0.2269	0.2037	0.1821	0.1698	0.1466	0.1173
	MRR@20	0.3188	0.3297	0.3367	0.3424	0.3332	0.3207	0.2954	0.2687	0.2446	0.2144	0.1830
SciDocs	Recall@10	0.1655	0.1705	0.1749	0.1776	0.1776	0.1753	0.1682	0.1603	0.1463	0.1295	0.1112
	nDCG@10	0.1578	0.1608	0.1652	0.1683	0.1680	0.1663	0.1591	0.1504	0.1356	0.1195	0.1031
	MAP@10	0.0923	0.0934	0.0959	0.0975	0.0970	0.0954	0.0904	0.0846	0.0747	0.0645	0.0545
	Precision@1	0.1810	0.1770	0.1870	0.1970	0.1920	0.1910	0.1810	0.1730	0.1480	0.1300	0.1160
	MRR@20	0.2864	0.2889	0.2982	0.3067	0.3061	0.3051	0.2946	0.2797	0.2527	0.2273	0.2031
Climate-FEVER	Recall@10	0.1750	0.1861	0.1947	0.1947	0.2016	0.2040	0.1952	0.1797	0.1651	0.1431	0.1235
	nDCG@10	0.1376	0.1473	0.1553	0.1592	0.1643	0.1651	0.1573	0.1464	0.1290	0.1098	0.0909
	MAP@10	0.0926	0.0996	0.1055	0.1100	0.1131	0.1134	0.1075	0.1010	0.0873	0.0737	0.0597
	Precision@1	0.1160	0.1270	0.1375	0.1433	0.1511	0.1498	0.1375	0.1290	0.1023	0.0834	0.0606
	MRR@20	0.1987	0.2120	0.2233	0.2310	0.2380	0.2376	0.2263	0.2121	0.1836	0.1558	0.1281
HotpotQA	Recall@10	0.6039	0.6199	0.6334	0.6419	0.6419	0.6318	0.6043	0.5615	0.5094	0.4534	0.4011
	nDCG@10	0.5741	0.5900	0.6022	0.6096	0.6090	0.5960	0.5679	0.5210	0.4651	0.4069	0.3531
	MAP@10	0.4800	0.4963	0.5080	0.5146	0.5131	0.4986	0.4702	0.4240	0.3724	0.3198	0.2725
	Precision@1	0.6663	0.6825	0.6951	0.7025	0.7020	0.6835	0.6509	0.5907	0.5163	0.4420	0.3749
	MRR@20	0.7427	0.7575	0.7692	0.7769	0.7771	0.7636	0.7360	0.6826	0.6142	0.5420	0.4736

Table 5: Sparse–dense hybridization results on Contriever.

Dataset	Metrics \ Ratio	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FiQA	Recall@10	0.3135	0.3311	0.3350	0.3404	0.3328	0.3320	0.3168	0.2857	0.2524	0.2292	0.2085
	nDCG@10	0.2500	0.2614	0.2673	0.2726	0.2699	0.2582	0.2378	0.2116	0.1880	0.1667	0.1496
	MAP@10	0.1895	0.1979	0.2031	0.2079	0.2065	0.1934	0.1750	0.1543	0.1372	0.1200	0.1066
	Precision@1	0.2299	0.2346	0.2392	0.2423	0.2361	0.2130	0.1914	0.1620	0.1451	0.1265	0.1127
	MRR@20	0.3188	0.3284	0.3366	0.3429	0.3402	0.3198	0.2906	0.2583	0.2326	0.2070	0.1865
SciDocs	Recall@10	0.1655	0.1701	0.1728	0.1748	0.1754	0.1757	0.1731	0.1696	0.1653	0.1582	0.1497
	nDCG@10	0.1578	0.1614	0.1652	0.1666	0.1666	0.1662	0.1632	0.1593	0.1532	0.1449	0.1370
	MAP@10	0.0923	0.0940	0.0964	0.0971	0.0967	0.0963	0.0940	0.0914	0.0866	0.0811	0.0765
	Precision@1	0.1810	0.1810	0.1880	0.1870	0.1860	0.1870	0.1780	0.1780	0.1690	0.1540	0.1460
	MRR@20	0.2864	0.2911	0.3001	0.3010	0.3007	0.3002	0.2957	0.2891	0.2779	0.2620	0.2498
Climate-FEVER	Recall@10	0.1750	0.1964	0.2196	0.2383	0.2620	0.2771	0.2868	0.2886	0.2814	0.2719	0.2598
	nDCG@10	0.1376	0.1555	0.1759	0.1937	0.2141	0.2271	0.2346	0.2346	0.2268	0.2170	0.2053
	MAP@10	0.0926	0.1052	0.1206	0.1343	0.1495	0.1584	0.1634	0.1627	0.1556	0.1475	0.1382
	Precision@1	0.1160	0.1355	0.1590	0.1759	0.1987	0.2046	0.2117	0.2104	0.1980	0.1811	0.1720
	MRR@20	0.1987	0.2227	0.2497	0.2741	0.3002	0.3173	0.3268	0.3258	0.3161	0.3019	0.2879
HotpotQA	Recall@10	0.6039	0.6238	0.6467	0.6663	0.6746	0.6654	0.6211	0.5658	0.5115	0.4612	0.4151
	nDCG@10	0.5741	0.5932	0.6121	0.6285	0.6342	0.6170	0.5625	0.5042	0.4495	0.4012	0.3605
	MAP@10	0.4800	0.4999	0.5187	0.5349	0.5406	0.5226	0.4706	0.4171	0.3669	0.3237	0.2877
	Precision@1	0.6663	0.6840	0.6980	0.7134	0.7132	0.6768	0.5951	0.5210	0.4550	0.4022	0.3625
	MRR@20	0.7427	0.7591	0.7728	0.7857	0.7878	0.7619	0.6906	0.6189	0.5540	0.4981	0.4542

Table 6: Query-format hybridization under the sparse retriever.

C Combinatorial Effect

To further examine how retriever-type and query-format hybridizations interact, we extend the analysis from § 3.4 by exploring their joint optimization space on the remaining datasets (Climate-FEVER, SciDocs, and HotpotQA). For each query, we perform an exhaustive grid search over three parameters—α (sparse-dense weighting), β_1 (original-expanded weighting for the sparse retriever), and β_2 (for the dense retriever)—to identify the configuration that maximizes nDCG@10.

Figures 10–12 visualize the distributions of query-wise optimal combinations across datasets. The results reveal strong interdependence between the two perspectives: the optimal sparse-dense ratio shifts depending on how the original and expanded queries are weighted, and vice versa. This suggests that each axis modulates the effectiveness of the other, reinforcing the necessity of a unified dual-perspective adaptive approach.

Dataset	Metrics \ Ratio	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FiQA	Recall@10	0.1649	0.1731	0.1779	0.1754	0.1810	0.1808	0.1719	0.1580	0.1541	0.1439	0.1304
	nDCG@10	0.1299	0.1350	0.1378	0.1372	0.1402	0.1406	0.1310	0.1209	0.1162	0.1053	0.0940
	MAP@10	0.0916	0.0949	0.0969	0.0975	0.0997	0.1017	0.0929	0.0847	0.0807	0.0724	0.0642
	Precision@1	0.1173	0.1173	0.1235	0.1265	0.1250	0.1219	0.1049	0.0957	0.0910	0.0787	0.0679
	MRR@20	0.1830	0.1870	0.1919	0.1916	0.1931	0.1892	0.1753	0.1642	0.1558	0.1388	0.1249
SciDocs	Recall@10	0.1112	0.1202	0.1307	0.1392	0.1449	0.1479	0.1465	0.1439	0.1435	0.1380	0.1322
	nDCG@10	0.1031	0.1115	0.1207	0.1296	0.1355	0.1408	0.1413	0.1401	0.1373	0.1313	0.1262
	MAP@10	0.0545	0.0598	0.0656	0.0718	0.0756	0.0799	0.0810	0.0806	0.0785	0.0747	0.0719
	Precision@1	0.1160	0.1240	0.1370	0.1470	0.1560	0.1710	0.1750	0.1740	0.1630	0.1500	0.1480
	MRR@20	0.2031	0.2161	0.2296	0.2448	0.2547	0.2668	0.2698	0.2689	0.2589	0.2465	0.2398
Climate-FEVER	Recall@10	0.1235	0.1365	0.1440	0.1491	0.1527	0.1486	0.1476	0.1404	0.1317	0.1245	0.1178
	nDCG@10	0.0909	0.1009	0.1089	0.1165	0.1195	0.1185	0.1184	0.1138	0.1065	0.1003	0.0942
	MAP@10	0.0597	0.0665	0.0721	0.0782	0.0795	0.0792	0.0792	0.0761	0.0707	0.0662	0.0614
	Precision@1	0.0606	0.0704	0.0821	0.0958	0.0977	0.1016	0.1062	0.1075	0.0990	0.0938	0.0879
	MRR@20	0.1281	0.1408	0.1546	0.1690	0.1737	0.1760	0.1768	0.1727	0.1634	0.1551	0.1469
HotpotQA	Recall@10	0.4011	0.4363	0.4753	0.5128	0.5411	0.5570	0.5470	0.5276	0.5041	0.4792	0.4560
	nDCG@10	0.3531	0.3860	0.4234	0.4596	0.4880	0.5013	0.4937	0.4770	0.4556	0.4339	0.4117
	MAP@10	0.2725	0.3010	0.3343	0.3679	0.3957	0.4102	0.4062	0.3934	0.3747	0.3557	0.3356
	Precision@1	0.3749	0.4097	0.4531	0.4928	0.5236	0.5357	0.5282	0.5091	0.4855	0.4652	0.4397
	MRR@20	0.4736	0.5120	0.5553	0.5947	0.6238	0.6316	0.6198	0.5981	0.5737	0.5502	0.5246

Table 7: Query-format hybridization under the dense retriever.

Dataset	**Metrics \ Methods**	**Individual Retriever**				**Single-Perspective**						**Dual-Perspective**	
						Retriever-Type		**Query-Form**					
		OS	OD	ES	ED	Static UB	Dynamic UB	Static UB (S)	Dynamic UB (S)	Static UB (D)	Dynamic UB (D)	Static UB	Dynamic UB
FiQA	Recall@10	0.3135	0.1649	0.2085	0.1835	0.3361	0.3916	0.3404	0.4007	0.1810	0.2278	0.3715	0.5011
	nDCG@10	0.2500	0.1299	0.1496	0.1363	0.2718	0.3272	0.2726	0.3341	0.1406	0.1851	0.2915	0.4385
	MAP@10	0.1895	0.0916	0.1066	0.0969	0.2087	0.2558	0.2079	0.2623	0.1017	0.1356	0.2209	0.3549
	Precision@1	0.2299	0.1173	0.1127	0.1049	0.2500	0.3179	0.2423	0.3287	0.1265	0.1744	0.2562	0.4522
	MRR@20	0.3188	0.1830	0.1865	0.1757	0.3424	0.4126	0.3429	0.4168	0.1931	0.2495	0.3592	0.5434
SciDocs	Recall@10	0.1655	0.1112	0.1497	0.1487	0.1776	0.2082	0.1757	0.2132	0.1479	0.1778	0.1917	0.2745
	nDCG@10	0.1578	0.1031	0.1370	0.1390	0.1683	0.2057	0.1666	0.2099	0.1413	0.1758	0.1813	0.2808
	MAP@10	0.0923	0.0545	0.0765	0.0783	0.0975	0.1226	0.0971	0.1261	0.0810	0.1032	0.1056	0.1772
	Precision@1	0.1810	0.1160	0.1460	0.1570	0.1970	0.2640	0.1880	0.2680	0.1750	0.2300	0.2060	0.3990
	MRR@20	0.2864	0.2031	0.2498	0.2584	0.3067	0.3750	0.3010	0.3775	0.2698	0.3273	0.3249	0.5015
Climate-FEVER	Recall@10	0.1750	0.1235	0.2598	0.1960	0.2040	0.2420	0.2886	0.3350	0.1527	0.1844	0.2972	0.3859
	nDCG@10	0.1376	0.0909	0.2053	0.1624	0.1651	0.2042	0.2346	0.2908	0.1195	0.1567	0.2408	0.3544
	MAP@10	0.0926	0.0597	0.1382	0.1121	0.1134	0.1426	0.1634	0.2085	0.0795	0.1084	0.1675	0.2613
	Precision@1	0.1160	0.0606	0.1720	0.1564	0.1498	0.2059	0.2117	0.3023	0.1075	0.1577	0.2150	0.4150
	MRR@20	0.1987	0.1281	0.2879	0.2393	0.2376	0.2940	0.3268	0.4113	0.1768	0.2335	0.3317	0.5136
HotpotQA	Recall@10	0.6039	0.4011	0.4151	0.4838	0.6419	0.6759	0.6746	0.7214	0.5570	0.6105	0.7110	0.7891
	nDCG@10	0.5741	0.3531	0.3605	0.4375	0.6096	0.6549	0.6342	0.6993	0.5013	0.5786	0.6694	0.7784
	MAP@10	0.4800	0.2725	0.2877	0.3592	0.5146	0.5630	0.5406	0.6149	0.4102	0.4916	0.5775	0.7075
	Precision@1	0.6663	0.3749	0.3625	0.4675	0.7025	0.7710	0.7134	0.8061	0.5357	0.6586	0.7463	0.8917
	MRR@20	0.7427	0.4736	0.4542	0.5527	0.7769	0.8288	0.7878	0.8557	0.6316	0.7282	0.8175	0.9198

Table 8: Static and dynamic upper bounds of hybridization.

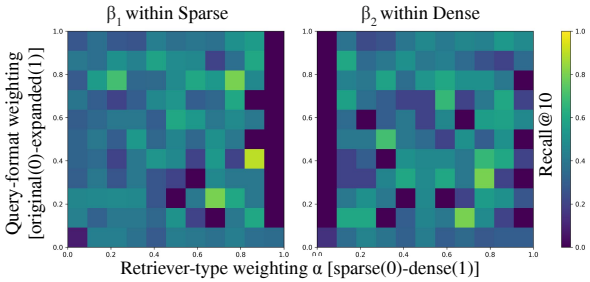

Figure 10: Heatmaps of per-query optimal weights in Scidocs.

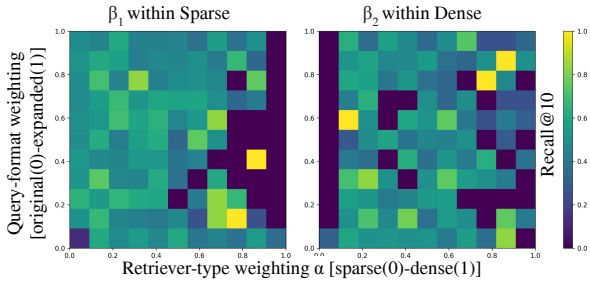

Figure 11: Heatmaps of per-query optimal weights in Climate-FEVER.

D Full Prompt of Scoring LLM

The following prompt was used to evaluate the relevance of top-1 retrieved passages for each query, as described in § 4. The LLM (`gpt-4o-mini`) receives the query and four candidate passages—one from each retrieval configuration—and assigns a score from 0 to 5 to indicate how well each passage answers the query. Passages are randomly shuffled before presentation to prevent order bias, and scores are mapped back to their corresponding retrievers after evaluation.

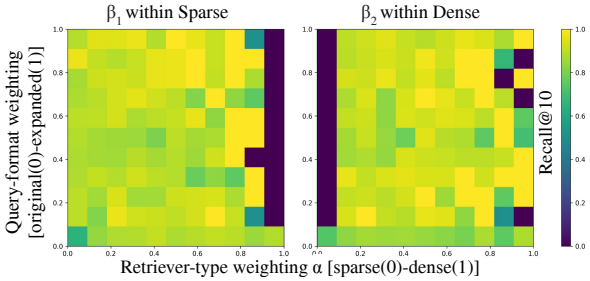

Figure 12: Heatmaps of per-query optimal weights in HotpotQA.

Prompt 1. Scoring LLM

You are an impartial evaluator of retrieval effectiveness. Score each anonymous passage with an integer from 0 to 5 (inclusive) using ONLY the criteria below. Return a JSON array of four integers in order, nothing else.

Scoring Criteria (0-5):
5 = Direct hit (clearly answers the question)
4 = Very close (high likelihood correct answer is nearby
3 = Somewhat close (related, partial answer or near-miss)
2 = Loosely related but likely off in ranking neighborhood
1 = Barely related; unlikely nearby
0 = Unrelated / off-track

Output: A JSON array of four integers (e.g., [3,4,5,1]). No extra text.

Query : `{Query}`

You will see 4 anonymous retrieval results (Top1s), in random order.
1 : `{Passage1}`
2 : `{Passage2}`
3 : `{Passage3}`
4 : `{Passage4}`

E Dataset Statistics and Details

To evaluate the performance and generalizability of our proposed method, we conduct experiments on five diverse datasets, including four from the BEIR benchmark (Thakur et al., 2021), FiQA, SciDocs, HotpotQA, and Climate-FEVER. These datasets span a variety of domains, query types, and semantic challenges, making them suitable for assessing the overall effectiveness of our method. Table 9 provides a summary of these datasets.

In particular, the four BEIR datasets often contain multiple relevant passages per query, making them ideal for comparing relative retrieval performance rather than performing strict single-document evaluation. Additionally, *HotpotQA* requires multi-hop reasoning, while *FiQA* and *SciDocs* are expert-domain datasets that demand domain-specific knowledge and reasoning.

- **FiQA** focuses on financial and economic question answering. Queries are fact-based and often domain-specific, requiring accurate terminology and high-precision retrieval. This dataset is well-suited to evaluate retrieval models in expert-oriented settings.
- **SciDocs** contains scientific article retrieval tasks such as citation prediction and document recommendation. Queries are typically paper abstracts or titles, and relevant documents may exhibit weak lexical overlap, posing a strong challenge for semantic retrieval.
- **HotpotQA** is an open-domain QA dataset designed for multi-hop reasoning. Each question requires evidence from multiple documents, making it useful for evaluating a system's ability to retrieve semantically complementary information.
- **Climate-FEVER** consists of fact-based claims related to climate change. The retrieval task is to find supporting or refuting evidence from scientific and journalistic corpora. This dataset emphasizes high-precision retrieval in a narrow but fact-intensive domain.

F Generalization Across Retrieval Components

To assess the generalizability of our findings, we conduct additional experiments using a broader set of retrieval components, including alternative sparse retrievers, dense retrievers, and query expansion methods. Specifically, we consider BM25 (Robertson and Walker, 1994) and SPLADE++ (Formal et al., 2021) as sparse retrievers, Contriever (Izacard et al., 2021), E5-base (Wang et al., 2022), and BGE-M3 (Chen et al., 2024) as dense retrievers, and two query expansion methods: HyDE-style self-generated passage (Gao et al., 2023; Zhang et al., 2024)s

	Domain	Task	# query	# corpus	average # gt/query
Scidocs	Pubmed paper	citation prediction	1,000	25K	4.9
FiQA	financial	QA	648	57K	2.6
Climate-Fever	climate wikipedia	fact checking	1,535	5.42M	3
HotpotQA	wikipedia	QA (multihop)	7,405	5.23M	2

Table 9: Statistics of the four datasets used for the evaluation.

Dataset	Method	Recall@10	nDCG@10	Precision@1
FiQA	Best Individual Retriever (OS)	0.3135	0.2500	0.2299
	Best Single-perspective (OS+OD)	0.3507	0.2797	0.2531
	QuDAR	**0.4082**	**0.3339**	**0.3086**
SciDocs	Best Individual Retriever (OS)	0.1655	0.1578	0.1810
	Best Single-perspective (OS+OD)	0.1837	0.1747	0.2010
	QuDAR	**0.1979**	**0.1883**	**0.2210**
Climate-FEVER	Best Individual Retriever (ED)	0.3234	0.2710	0.2580
	Best Single-perspective (OD+ED)	0.3641	0.3045	0.2951
	QuDAR	**0.3777**	**0.3208**	**0.3147**
HotpotQA	Best Individual Retriever (OS)	0.6039	0.5741	0.6663
	Best Single-perspective (OS+OD)	0.6883	0.6562	0.7473
	QuDAR	**0.7417**	**0.7001**	**0.7646**

Table 10: Generalization results with BM25 (sparse), BGE-M3 (dense), and HyDE query expansion. Best results are in bold.

and pseudo-relevance feedback (PRF) (Rocchio Jr, 1971). These configurations cover a range of retrieval architectures and expansion strategies commonly used in recent work.

Tables 10–13 summarize representative results under these settings. Across all configurations, we observe that the patterns identified in § 5 remain consistent. First, substantial per-query variability persists, indicating that the optimal combination of retriever type and query format varies across queries even when stronger or more recent components are used. Second, the two perspectives exhibit structured interactions, consistent with the analysis in § 3.3–§ 3.4. This confirms that dual-perspective interaction is not tied to a specific model or method, but reflects a general property of retrieval behavior.

In terms of performance, QUDAR remains competitive with or outperforms individual retrievers and single-perspective methods across most configurations. The largest gains are observed when neither retriever-type nor query-format signals are dominant, i.e., when complementary signals can be effectively combined. In contrast, when a single signal consistently dominates, single-perspective methods may achieve stronger performance. This suggests that the effectiveness of QUDAR depends on the degree of complementarity between the two perspectives. Overall, these results indicate that the dual-perspective structure identified in this work is largely model-agnostic and generalizes across a wide range of retrieval architectures.

G Full Prompt of Query Exapnsion

We employ `LLaMA 3.1-8B` to generate expanded queries by reformulating the original query into a self-contained passage, following prior work that demonstrated its effectiveness for improving recall and semantic coverage (Gao et al., 2023; Zhang et al., 2024). The following prompt was used for all datasets and retrievers.

Prompt 2. Query Expansion

You are a neutral fact compiler. Your task is to write a single, dense, factual *paragraph* with no formatting. Crucially, do not answer the query. Your goal is to write a *descriptive passage* about the topic or entity that the answer refers to. Within this description, you must naturally embed the specific details mentioned in the query. The final connection must be inferred by the reader.

Query : `{Query}`
Passage :

Dataset	Method	Recall@10	nDCG@10	Precision@1
FiQA	Best Individual Retriever (OS)	0.3135	0.2500	0.2299
	Best Single-perspective (OS+OD)	0.3507	0.2797	0.2531
	QuDAR	**0.3629**	**0.2958**	**0.2654**
SciDocs	Best Individual Retriever (OS)	0.1655	0.1578	0.1810
	Best Single-perspective (OS+OD)	0.1837	0.1747	**0.2010**
	QuDAR	**0.1948**	**0.1783**	0.1800
Climate-FEVER	Best Individual Retriever (OD)	0.3013	0.2501	0.2293
	Best Single-perspective (OD+ED)	0.3255	**0.2676**	**0.2410**
	QuDAR	**0.3256**	0.2572	0.1987
HotpotQA	Best Individual Retriever (OS)	0.6039	0.5741	0.6663
	Best Single-perspective (OS+OD)	**0.6883**	**0.6562**	**0.7473**
	QuDAR	0.6637	0.6298	0.7209

Table 11: Generalization results with BM25 (sparse), BGE-M3 (dense), and PRF query expansion. Best results are in bold.

Dataset	Method	Recall@10	nDCG@10	Precision@1
FiQA	Best Individual Retriever (OD)	0.4781	0.4065	0.3981
	Best Single-perspective (OS+OD)	**0.5122**	**0.4349**	**0.4228**
	QuDAR	0.4719	0.3859	0.3457
SciDocs	Best Individual Retriever (OD)	0.1995	0.1847	0.2020
	Best Single-perspective (OD+ED)	**0.2057**	0.1930	0.2220
	QuDAR	0.2044	**0.1951**	**0.2330**
Climate-FEVER	Best Individual Retriever (OD)	0.3209	0.2638	0.2391
	Best Single-perspective (OD+ED)	0.3602	0.3016	0.2853
	QuDAR	**0.3637**	**0.3074**	**0.3003**
HotpotQA	Best Individual Retriever (OD)	0.7292	0.6910	0.7623
	Best Single-perspective (OS+OD)	0.7640	0.7294	0.8120
	QuDAR	**0.7657**	**0.7334**	**0.8205**

Table 12: Generalization results with BM25 (sparse), E5-base (dense), and HyDE query expansion. Best results are in bold.

H Effect of LLM-based Weighting

To isolate the effect of LLM-based weighting, we compare its performance under single-perspective settings (retriever-type only and query-format only) and the dual-perspective setting used in QUDAR. Specifically, we apply the same LLM-based scoring mechanism across these variants and evaluate their retrieval performance.

Table 14 shows that LLM-based weighting consistently improves over individual retrievers, indicating that LLM-derived relevance signals are effective. However, the dual-perspective variant (QUDAR-llm) achieves higher performance than both single-perspective variants in most cases. This suggests that the performance gains are not solely due to the LLM itself, but also depend on combining complementary signals across retriever-type and query-format.

We also observe that when one signal is particularly strong, single-perspective variants can be competitive or even outperform the dual-perspective approach. For example, on CLIMATE-FEVER, the query-format variant slightly exceeds QUDAR-llm, reflecting a case where expanded queries with sparse retrieval dominate. Overall, these results indicate that the effectiveness of QUDAR-llm depends on the degree of complementarity between the two perspectives.

Dataset	Method	Recall@10	nDCG@10	Precision@1
FiQA	Best Individual Retriever (OS)	0.4236	0.3568	0.3395
	Best Single-perspective (OS+ES)	**0.4299**	**0.3630**	**0.3549**
	QuDAR	0.3971	0.3330	0.3133
SciDocs	Best Individual Retriever (OS)	0.1641	0.1581	0.1910
	Best Single-perspective (OS+OD)	0.1804	0.1701	0.2050
	QuDAR	**0.1919**	**0.1815**	**0.2090**
Climate-FEVER	Best Individual Retriever (ES)	0.2974	0.2496	0.2391
	Best Single-perspective (OS+ES)	**0.3472**	**0.2914**	**0.2801**
	QuDAR	0.3073	0.2572	0.2463
HotpotQA	Best Individual Retriever (OS)	0.7023	0.6867	0.8131
	Best Single-perspective (OS+ES)	**0.7398**	**0.7166**	**0.8334**
	QuDAR	0.7339	0.7069	0.8054

Table 13: Generalization results with SPLADE++ (sparse), Contriever (dense), and HyDE query expansion. Best results are in bold.

Dataset	Method	Recall@10	nDCG@10	Precision@1
FiQA	Single-perspective: Retriever-type	0.3230	0.2656	**0.2485**
	Single-perspective: Query-format	0.3271	0.2604	0.2222
	Dual-perspective: QUDAR-llm	**0.3639**	**0.2837**	0.2454
SciDocs	Single-perspective: Retriever-type	0.1737	0.1640	0.1900
	Single-perspective: Query-format	0.1770	0.1684	0.1920
	Dual-perspective: QUDAR-llm	**0.1920**	**0.1816**	**0.2050**
Climate-FEVER	Single-perspective: Retriever-type	0.2019	0.1646	0.1505
	Single-perspective: Query-format	**0.2803**	**0.2311**	**0.2137**
	Dual-perspective: QUDAR-llm	0.2583	0.2138	0.2085
HotpotQA	Single-perspective: Retriever-type	0.6357	0.6099	0.7184
	Single-perspective: Query-format	0.6681	0.6331	0.7219
	Dual-perspective: QUDAR-llm	**0.6883**	**0.6579**	**0.7515**

Table 14: Comparison of LLM-based weighting under single-perspective and dual-perspective (QUDAR-llm) settings. Best results are in bold.