A Survey of Reasoning-Intensive Retrieval: Progress and Challenges

Yiyang Wei[1*] **Tingyu Song**[2*] **Siyue Zhang**[3] **Yilun Zhao**[4]
[1]Zhejiang University [2]University of the Chinese Academy of Sciences
[3]Nanyang Technological University [4]Yale University

Abstract

Reasoning-Intensive Retrieval (RIR) targets retrieval settings where relevance is mediated by latent inferential links between a query and supporting evidence, rather than semantic similarity. Motivated by the emergent reasoning abilities of Large Language Models (LLMs), recent work integrates these capabilities into the IR field, spanning the entire pipeline from benchmarks to retrievers and rerankers. Despite this progress, the field lacks a systematic framework to organize current efforts and articulate a clear path forward. To provide a clear roadmap for this rapidly growing yet fragmented area, this survey (1) systematizes existing RIR benchmarks by knowledge domains and modalities, providing a detailed analysis of the current landscape; (2) introduces a structured taxonomy that categorizes methods based on where and how reasoning is integrated into the retrieval pipeline, alongside an analysis of their tradeoffs and practical applications; and (3) summarizes challenges and future directions to guide research in this evolving field.

1 Introduction

Information Retrieval (IR) underpins everyday information access (*i.e.,* web search) and has advanced rapidly in real world applications (Devlin et al., 2019; Izacard et al., 2022). Within the rise of deep research and agentic search (Qiao et al., 2025; Shi et al., 2025), retrieval has increasingly extended to more scenarios such as multi-hop (Yang et al., 2018), instruction-following (Weller et al., 2025a,b), and long-context retrieval (Zhu et al., 2024; Saad-Falcon et al., 2024).

These advances aim for scenarios with high semantic overlap. However, retrieval in expert domains requires not just overcoming lexical or semantic distances, but a deeper reasoning capability

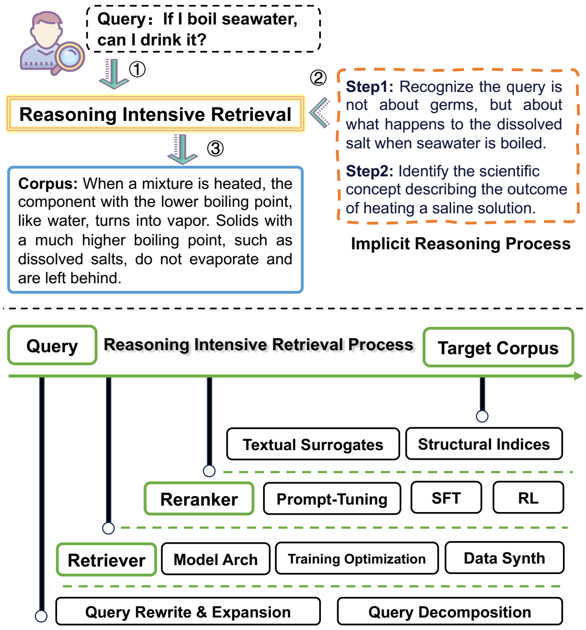

Figure 1: **Top:** An example of reasoning-intensive retrieval, where a query and its supporting document are connected through an implicit multi-hop reasoning chain. **Down:** Overview of the retrieval pipeline and representative techniques, which is detailed in Section 4.

to infer implicit connections, such as mapping a brief algorithm description to its symbolic code. We refer to this setting as **R**easoning- **I**ntensive **R**etrieval (**RIR**) where relevance is based on latent inferential links connecting a query to supporting evidence. For example, as shown in Figure 1, answering whether boiled seawater is drinkable requires retrieving evidence about the behavior of dissolved salt during boiling, even though the query and the relevant document are linked only through an implicit multi-hop reasoning chain rather than direct lexical overlap.

To evaluate the corresponding abilities of current retrieval systems, BRIGHT is introduced as an early benchmark (Su et al., 2025). Subsequent efforts have extended RIR evaluation to domain-specific scenarios (Zheng et al., 2025; Li et al.,

* Equal contributions. Correspondence to: Tingyu Song (songtingyu23@mails.ucas.ac.cn), Yilun Zhao (yilun.zhao@yale.edu)

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 42101–42122
July 2-7, 2026 ©2026 Association for Computational Linguistics

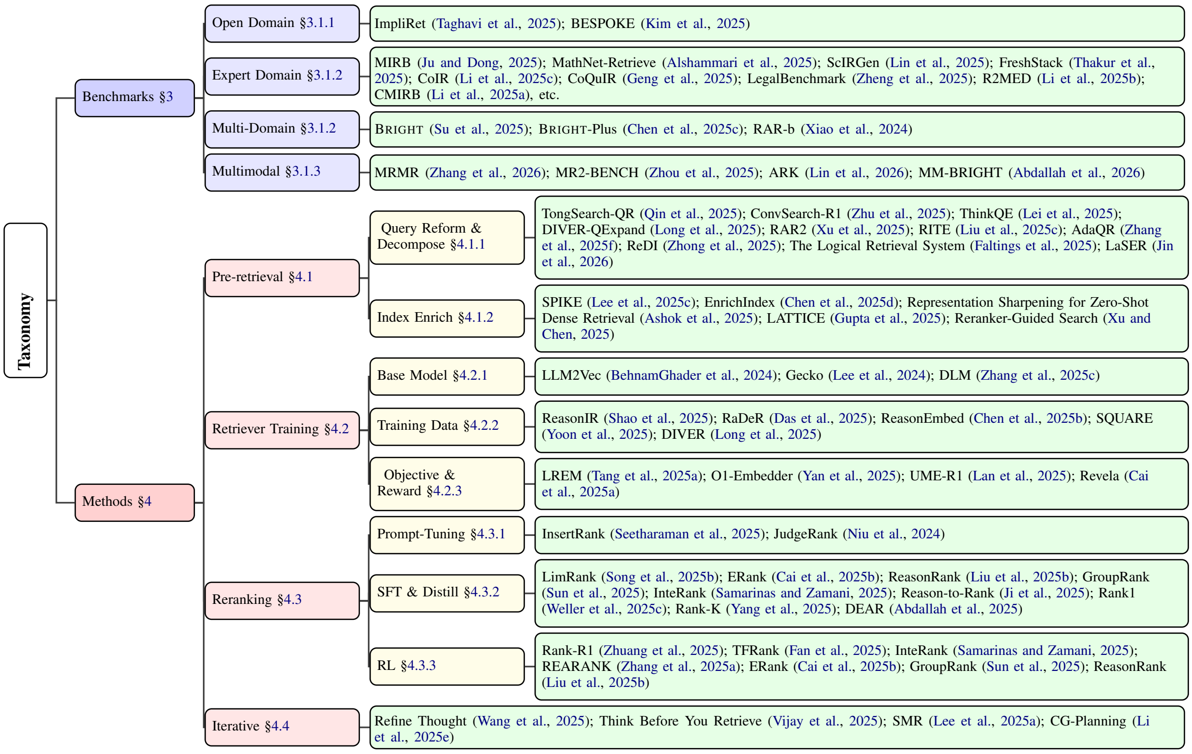

Figure 2: Taxonomy of Reasoning-Intensive Retrieval (RIR).

2025b; Ju and Dong, 2025) and multimodal settings (Zhang et al., 2026; Zhou et al., 2025), exposing the limitations of state-of-the-art retrievers. Motivated by these findings, a growing family of methods integrate reasoning into different stages of the retrieval pipeline, through query-side transformation (Qin et al., 2025; Lei et al., 2025; Xu et al., 2025), reasoning-aware representation learning (Shao et al., 2025; Long et al., 2025; Lan et al., 2025), and reranking (Song et al., 2025b; Zhuang et al., 2025; Liu et al., 2025b), to improve retrieval performance on reasoning-intensive queries. Recent studies further suggest that effective RIR may require iterative retrieval pipelines that repeatedly alternate between retrieval and reasoning (Wang et al., 2025; Vijay et al., 2025).

Despite this rapid progress, existing RIR research still faces two main limitations. First, the evaluation landscape remains highly heterogeneous. Current studies adopt diverse problem formulations, datasets, and evaluation setups across tasks and domains (*e.g.,* code, biomedical, math). Second, methodological developments are scattered across different stages of the retrieval pipeline, including query rewriting, retriever training, reranking, and iterative retrieval frameworks. As a result, the field remains difficult to navigate and lacks consistent evaluation and methodological organization. In this survey, we aim to address these issues by (1) systematizing existing benchmarks according to reasoning type, domain, and source of difficulty (Section 3); (2) proposing a structured taxonomy of RIR methods based on where reasoning is introduced in the retrieval pipeline (Section 4), and analyzing their trade-offs and application scenarios (Appendix D); and (3) outlining key open challenges in evaluation metrics, domain generalization, inference cost, and multimodal reasoning (Section 5).

2 Related Work

Reasoning-intensive Retrieval (RIR) is a nascent but rapidly emerging domain. However, to the best of our knowledge, comprehensive surveys of this field are still scarce. Existing IR surveys have made substantial contributions in cataloguing the evolution of retrieval paradigms (Robertson and Zaragoza, 2009; Yates et al., 2021; Li et al., 2025d; Zhang et al., 2025b), but these works primarily focus on semantic or lexical query-document relevance, leaving the inferential demands placed on the retrieval system largely unaddressed. When it comes to the intersection between reasoning and retrieval, current surveys often emphasize the role of reasoning within RAG and agentic frameworks, such as RAG-Reasoning (Li et al., 2025g) and Rea-

Domain	Name	Size	Annotation Type
Open Domain	ImpliRet	9,000	LLM-Automated
	BESPOKE	150	Human-Curated
Scientific	MIRB	39,029	Derived[1]
	MathNet-Retrieve	10,000	Hybrid[2]
	SciRGen	61,376	LLM-Automated
	FreshStack	672	LLM-Automated
Code	CoIR	≈162,000	Derived
	CoQuiR	42,725	LLM-Automated
Legal	Legal-Benchmark	9,863	Human-Curated
Medical	R2MED	876	Hybrid
	CMIRB	10,962	LLM-Automated
Multi-Domain	BRIGHT	1,384	Hybrid
	BRIGHT-Plus	1,384	Hybrid
	RAR-b	45,745	Derived
Multi-Modal	MRMR	1,435	Hybrid
	MR2-BENCH	1,309	Hybrid
	ARK	1,547	Hybrid
	MM-BRIGHT	2,803	Hybrid

Table 1: Summary of RIR evaluation Benchmarks (see full table in Table 3 in Appendix). [1]Derived: Source is derived from established data sources (*e.g.,* previous datasets, libraries, internet QA). [2]Hybrid: Source is both LLM-Automated and Human-Curated.

soning Agentic RAG (Liang et al., 2025), they typically treat retrieval as a preliminary stage to support generation. These works prioritize how to leverage retrieved evidence for reliable answers rather than the inferential depth of the retrieval process itself. In contrast, RIR focuses on the retrieval system's intrinsic ability to infer connections between a query and the target corpus through implicit logical inferential links (Su et al., 2025; Zhang et al., 2026). In this setting, retrieval is the end task under a framework of inference-mediated relevance.

3 Reasoning-Intensive IR Evaluation

In this section, we compile existing benchmarks for reasoning-intensive retrieval and provide a comparative analysis across them.

3.1 Existing Evaluation Benchmarks

Current reasoning-intensive retrieval benchmarks cover a broad range of domains. We classify them into the following four types: (1) *open-domain*, which covers general-purpose knowledge and commonsense reasoning; (2) *expert-domain*, which probes specialized knowledge within a single professional discipline; (3) *multi-domain*, which aggregates tasks from multiple professional areas to test knowledge breadth; (4) *multimodal*, which introduces unique challenges distinct from text-only processing and represents a significant frontier. We first provide a brief summary of these benchmarks in Table 1, and present a more comprehensive overview in Table 3 in the Appendix. We next detail these evaluation benchmarks:

3.1.1 Open-Domain Benchmarks

Open-domain benchmarks operate on general-purpose knowledge and commonsense, without requiring specialized expertise. The primary reasoning challenge in these daily settings is to decipher the user's latent intent, which is often implicit and context-dependent. To this end, the BESPOKE (Kim et al., 2025) and ImpliRet (Taghavi et al., 2025) benchmarks construct evaluation frameworks using user chat histories, where queries are frequently short and ambiguous. They pose a significant challenge to current models by explicitly testing their ability to recover underlying intent from the conversational context, providing a realistic measure of current models' practical utility.

3.1.2 Expert-Domain Benchmarks

Expert-domain benchmarks address professional fields where specialized knowledge and domain-specific practices complicate relevance assessment, necessitating reasoning abilities beyond what is required in general settings.

Scientific. The scientific domain encompasses fields built on formal systems of knowledge. For instance, ScIRGen (Lin et al., 2025) addresses the lack of realism in scientific QA benchmarks by proposing a scalable generation framework that creates complex, task-implicit questions grounded in papers. FreshStack (Thakur et al., 2025) is the first to deliver an automated retrieval evaluation benchmark tailored to real developer needs in technical documentation domain. In mathematics, MIRB (Ju and Dong, 2025) and MathNet-Retrieve (Alshammari et al., 2025) evaluate whether systems can retrieve mathematically relevant statements. While MathNet-Retrieve (Alshammari et al., 2025) focuses on equivalent problems across multilingual and multimodal contexts, MIRB(Ju and Dong, 2025) extends the evaluation to more reasoning tasks, including theorem-level premise retrieval and problem-solving answer retrieval.

Legal. Legal retrieval is challenging because it requires bridging abstract legal rules with concrete, case-specific situations. This challenge extends to precedent retrieval, which involves identifying legally analogous cases that share overlapping legal principles (Nigam et al., 2022; Li et al., 2023). To evaluate this reasoning capability directly, a new benchmark (Zheng et al., 2025) introduces two reasoning-intensive tasks, Bar Exam QA and Housing Statute QA, which require systems to connect factual scenarios to their governing statutes through analytical and deductive reasoning.

Medical. In medicine, a similar challenge arises, but the ambiguity stems not from abstract rules but from underspecified, symptom-centered queries. Benchmarks like R2MED (Li et al., 2025b) and CMIRB (Li et al., 2025a) evaluate retrieval for vague patient presentations, where relevance is determined by linking symptoms to plausible diagnoses and appropriate treatment plans.

Code. Compared with natural-language retrieval, reasoning-intensive retrieval in code demands reasoning over symbols and structure. CoIR (Li et al., 2025c), for instance, assesses a model's ability to reason about program behavior through tasks like cross-language code equivalence and bug localization. Building on this, CoQuIR (Geng et al., 2025) pushes further by demanding that retrievers discriminate not only by functionality but also by code quality, with attributes including correctness, efficiency, and security. These benchmarks signal a shift from retrieving topically relevant code (Husain et al., 2019) to identifying high-quality, reliable solutions.

Multi-Domain Benchmarks. In contrast to benchmarks focused on a single domain, multi-domain benchmarks aggregate representative tasks from several professional fields to provide a broader evaluation of current models' capabilities. For example, BRIGHT (Su et al., 2025) and BRIGHT-Plus (Chen et al., 2025c) exemplify this direction by covering specialized areas such as science, technology, engineering, and mathematics, and by including queries on topics such as software debugging and scientific theorem retrieval. Meanwhile, RAR-b (Xiao et al., 2024) derives retrieval instances from multiple-choice QA to probe diverse reasoning skills (*e.g.,* commonsense, temporal), but its shorter retrieval targets make it closer to conceptual capability testing than document-level professional search.

3.1.3 Multimodal Benchmarks

Multimodal RIR benchmarks introduce novel challenges by moving beyond text-only retrieval to tasks that demand reasoning across diverse modalities (*e.g.,* image, text). Recent multimodal retrieval benchmarks, including MRMR (Zhang et al., 2026), MM-BRIGHT (Abdallah et al., 2026), and ARK (Lin et al., 2026), introduce reasoning-heavy and knowledge-intensive tasks that require models to capture abstract conceptual connections across scientific multimodal documents and diverse domains. In contrast, MR^2-Bench (Zhou et al., 2025) broadens the task scope and places stronger emphasis on evaluating spatial, logical, and causal reasoning capabilities through challenging scenarios such as visual puzzles and dimensional transformations.

3.2 Comparative Benchmark Analysis

Having reviewed the landscape of reasoning-intensive IR benchmarks across domains and modalities, we now turn to a comparative analysis that highlights two key axes: the scale-reliability trade-off in benchmark construction and emphasis on different reasoning types across domains.

Scale–Reliability Trade-offs. A fundamental trade-off exists in benchmark design between scalable synthetic generation and rigorous human curation (see Table 1 and Table 3). On one hand, LLM-based synthetic benchmarks like ScIRGen (Lin et al., 2025) and ImpliRet (Taghavi et al., 2025) expand coverage and diversify cognitive demands, but can suffer from hallucinations and limited validation. On the other hand, reliability-oriented benchmarks emphasize human/expert oversight, including BRIGHT (Su et al., 2025) and its cleaned extension BRIGHT-Plus (Chen et al., 2025c) sources data from human experts across various domains to ensure trustworthiness. This emphasis on reliability becomes paramount in high-stakes fields, such as Bar Exam QA (Zheng et al., 2025). Thus, an open direction is hybrid construction pipelines that scale via synthesis while preserving evaluative validity through targeted expert checks.

Reasoning Types and Domain Emphases. Following BRIGHT (Su et al., 2025), we categorize RIR benchmarks into five reasoning types—*deductive*, *analogical*, *causal*, *analytical*, and *numerical*, as summarized in Table 2. We provide representative examples and inference chains for each type in Table 4 in Appendix. *Numerical reasoning* often involves arithmetic or temporal operations in daily settings (Taghavi et al., 2025), whereas *deductive reasoning* is the most prevalent across domains, supporting rule-to-case application in math-

Deductive Reasoning: A general principle or theorem in the document is directly applied to explain a specific scenario or solve a problem in the query.
Analogical Reasoning: A document draws a parallel with the query in its underlying logic, indicating that the query and document share a solution strategy or a common theorem/algorithmic foundation.
Causal Reasoning: The document identifies root causes or mechanistic relationships that explain effects observed in the query. Resolution requires tracing causal chains from symptoms to origins.
Analytical Reasoning: The document provides critical domain knowledge that fills gaps in multi-step reasoning chains required to resolve the query. This involves decomposition of complex problems into interdependent sub-questions.
Numerical Reasoning: The query is resolved by applying quantitative constraints in the document, requiring arithmetic computation (*e.g.,* percentages, unit conversion, rate/ratio) or time arithmetic (*e.g.,* duration, scheduling offsets, temporal comparisons). The logical mechanism is a deterministic mapping from numeric facts and rules to a target value or decision.

Table 2: Definitions of the five reasoning types covered by existing RIR benchmarks.

ematics/science (Ju and Dong, 2025), medicine (Li et al., 2025b,a), and law (Zheng et al., 2025). *Analogical reasoning* is particularly salient in code (Li et al., 2025c; Su et al., 2025) and math (Alshammari et al., 2025; Ju and Dong, 2025) benchmarks for establishing functional correspondences across modalities. Finally, *causal* and *analytical* reasoning frequently appear in specialized tasks such as troubleshooting and problem decomposition.

4 Reasoning-Intensive IR Methods

Reasoning-intensive retrieval can inject reasoning at different points of the retrieval pipeline, from shaping the input to refining candidates during ranking and multi-step interaction. To make these design choices comparable, we organize existing methods by *where* reasoning is introduced and *how* it interacts with retrieval. Accordingly, we structure this section into four stages: pre-retrieval augmentation, retrieval, reranking, and iterative workflows. To complement this structural taxonomy, Appendix D provides a comparative analysis across these categories, while Appendix F maps the methods to specific downstream tasks and applications.

4.1 Pre-Retrieval Reasoning Augmentation

To enhance RIR, pre-processing techniques can be applied to both queries and documents before the matching stage. Query-side methods (§4.1.1) focus on refining or decomposing user's request to clarify its underlying intent; document-side methods (§4.1.2) aim to enrich the document corpus, making latent evidence more explicit and accessible.

4.1.1 Query-Side Augmentation

Query-side augmentation methods can be broadly grouped into the following two categories:

Query Rewriting and Expansion. Query rewriting and expansion leverages LLM-generated reasoning traces to reformulate or enrich the original query, aiming to make the underlying information need more explicit for downstream retrieval. TongSearch-QR (Qin et al., 2025) and ConvSearch-R1 (Zhu et al., 2025) leverage Reinforcement Learning (RL) with thinking format reward and performance reward to train LLM on query rewriting tasks, achieving better performance with smaller model size. In addition, ConvSearch-R1 (Zhu et al., 2025) adopts a cold-start supervised fine-tuning (SFT) stage before RL to improve output format adherence and stabilize reasoning and rewriting behaviors. For query expansion, RAR2 (Xu et al., 2025) fine-tunes LLMs with a thought dataset and Direct Preference Optimization (DPO) to generate reasoning traces that augment retrieval in clinical scenarios. Moving beyond a single-pass expander, ThinkQE (Lei et al., 2025) formulates query expansion as an interactive process that iteratively refines expansions using retrieval feedback, and DIVER-QExpand (Long et al., 2025) simplifies this workflow by retaining the original and the final-round expansion to control token growth while preserving key information. Beyond text-only rewriting, AdaQR (Zhang et al., 2025f) and LaSER (Jin et al., 2026) produce latent reasoning in the embedding space, increasing retrieval performance while maintaining low inference latency. Beyond single-vector retrieval, AMER (Chen et al., 2025a) autoregressively generates multiple query embeddings for retrieval, outperforming single-embedding baselines.

Query Decomposition. Query decomposition breaks a complex query into sub-queries to better capture multifaceted intents. This strategy is particularly relevant to *analytical reasoning* retrieval, where solving the task typically requires a multi-step reasoning chain, that each step can be operationalized as a sub-query. ReDI (Zhong et al., 2025) exemplifies this approach with a three-stage pipeline that performs intent recognition, enriches sub-queries for efficient parallel retrieval, and fuses the retrieved results, leveraging LLM reasoning throughout. In contrast, the logical retrieval system (Faltings et al., 2025) decomposes natural-language queries into sub-queries connected by logical operators (*e.g.,* OR, AND, NOT) and aggregates cosine-similarity signals to better handle compositional constraints.

4.1.2 Index-Side Augmentation

Complementing query rewriting, index-side augmentation shifts the reasoning burden to offline ingestion by pre-enriching document representations with synthetic metadata. We group existing index-side techniques into the following two types:

Textual Surrogates. Textual-surrogate methods expand each document with auxiliary descriptions that anticipate how users might seek it, while remaining compatible with standard dense retrieval pipelines. SPIKE (Lee et al., 2025c) instantiates this idea by generating hypothetical retrieval scenarios for each document. Similarly, representation sharpening (Ashok et al., 2025) strengthens index representations via *document-conditioned* contrastive queries that emphasize distinguishing aspects of a document. These methods expose the implicit information needs a document could satisfy, enhancing semantic coverage to better support reasoning-driven inferential links. Beyond effectiveness, EnrichIndex (Chen et al., 2025d) highlights a practical benefit of such enrichment: by shifting semantic expansion offline, enriched indices can reduce repeated online LLM computation during retrieval, lowering latency and cost.

Structural Indices. While textual surrogates improve final performance through additional views, structural indices externalize reasoning pathways by organizing knowledge into interpretable frameworks that retrieval can traverse. LATTICE (Gupta et al., 2025) exemplifies this direction by constructing LLM-guided lattice structures that enable multi-level navigation, capturing implicit dependencies and supporting complex reasoning queries through coarse-to-fine exploration. Similarly, reranker-guided search (Xu and Chen, 2025) couples retrieval with downstream ranking signals to steer exploration toward higher-utility regions of the corpus, effectively using structured search trajectories to refine retrieval decisions.

4.2 Reasoning-Aware Retriever Training

To improve the retrievers' reasoning performance in RIR domain, current efforts mainly focus on three aspects: (1) *model architecture selction*, current methods implement their algorithm to different architecture of models; (2) *data curation*, some works carefully curate training data specialized for RIR; and (3) *training objectives and reward design* used during optimization.

4.2.1 Base Model Architecture Selection

Different embedding backbone is a design choice for RIR. Motivated by the strong reasoning abilities of LLMs, several works (BehnamGhader et al., 2024; Lee et al., 2024) adapt decoder-style architectures for dense embedding models, yielding LLM-based retrievers. However, their unidirectional attention limits the effectiveness of incorporating bidirectional context. In contrast, Diffusion Language Model (DLM) embeddings (Zhang et al., 2025c) leverage bidirectional attention to better integrate surrounding information, which improves reasoning efficiency and embedding performance.

4.2.2 Training Data Curation

Despite the importance of the backbone, training data largely determines which reasoning patterns the model can actually learn to represent. Curating specialized training data infused with reasoning elements is an important strategy for boosting the performance of retrievers on logic-heavy queries.

To curate high quality documents for RIR, the central reasoning challenge is supervision mining, positive documents should provide evidence that genuinely supports answering the query (Yoon et al., 2025), while negatives should remain lexically or semantically similar to the query yet be unhelpful to resolve it (Shao et al., 2025; Long et al., 2025). For positives, SQUARE (Yoon et al., 2025) uses LLM-generated hypothetical answers to retrieve and verify supportive positives. And to curate hard negatives for RIR, ReasonIR (Shao et al., 2025) and DIVER (Long et al., 2025) perform iterative mining guided by LLM-generated rationales, ReasonEmbed (Chen et al., 2025b) further filters candidates using embedding models with LLM relevance annotations, and RaDeR (Das et al., 2025) leverages MCTS with an LLM to synthesize diverse hard-negative training signals.

In contrast, for (query, thought, document) triplets, where reasoning is realized through generating retrieval "thoughts", the central challenge is to synthesize and retain only those thoughts that provide genuine retrieval utility. For instance, O1 Embedder (Yan et al., 2025) addresses this by prompting an expert LLM to produce candidate thoughts and filtering them via a retrieval committee. On top of that, LREM (Tang et al., 2025a) curates training signals by comparing retrieval outcomes with and without the thought, discarding queries that yield no improvement.

4.2.3 Training Objectives and Reward Design

With reasoning-capable backbones and reasoning-intensive supervision in place, an important step is to choose objectives and rewards that internalize these signals into the retriever's embedding and ranking behaviors. A representative direction is multi-task optimization that jointly trains (1) reasoning generation and (2) embedding discrimination (details of loss functions and analysis are in Appendix E). For example, LREM and O1 Embedder (Tang et al., 2025a; Yan et al., 2025) combine next-token prediction over intermediate thoughts with contrastive losses, typically via a weighted sum, so that the model learns to "think" while remaining a competitive embedder. In contrast, the Dense Reasoner (Zhang et al., 2025f) distills the effect of LLM reasoning directly into the embedding space by learning an embedding transformation with an MSE objective that matches LLM-reasoned embeddings. Extending joint objectives to multimodal retrieval, UME-R1 (Lan et al., 2025) integrates discriminative contrastive learning with generative objectives defined over reasoning trajectories, together with next-token prediction during cold-start SFT, to support both discriminative and reasoning-driven generative embeddings across modalities. Revela (Cai et al., 2025a) optimizes the retriever directly via a language-modeling objective with in-batch attention, enabling self-supervised retriever learning without query–document pairs.

Building on the above strategies to enhance the RIR performance, RL-based alignment further makes the reasoning trajectory itself an explicit optimization target by shaping it with structured rewards. In LREM (Tang et al., 2025a), a RL stage scores sampled CoTs with a weighted combination of *generation-side* rewards (*e.g.,* format compliance and length control) to encourage structured yet concise trajectories, together with an *embedding-side* retrieval-accuracy reward that favors trajectories producing embeddings with stronger discriminative separation. Similarly, with both generation-side and embedding-side rewards, UME-R1 (Lan et al., 2025) grounds multimodal representation learning in reasoning trajectories, thereby steering training toward higher-quality reasoning-conditioned multimodal embeddings.

4.3 Reasoning-Enhanced Reranking

Given the retrieved documents, a reranker needs to refine documents' order by evaluating documents from multiple perspectives, which involves deeper reasoning abilities to surface the most useful evidence for the query. To clarify how rerankers acquire and strengthen such reasoning ability, we group existing approaches into three paradigms: (1) Prompt-Tuning, which is conducted during the inference time, (2) Supervised reasoning transfer, which is often realized by SFT and Distillation, and (3) Reinforcement Learning, which further improves the general abilities in RIR.

4.3.1 Prompt-Tuning

Prompted rerankers elicit reasoning at inference time without parameter updates, making them attractive for rapid deployment and out-of-domain transfer. InsertRank (Seetharaman et al., 2025) inserts BM25 score into the prompt to help reranker reasoning on relevance. In addition, JudgeRank (Niu et al., 2024) leverages agentic prompting further decomposing reranking into stages such as query analysis and document analysis, which improves robustness on corresponding queries.

4.3.2 Supervised Reasoning Transfer

While prompted rerankers largely rely on inference-time, supervised transfer aims to *internalize* reasoning behaviors through training on curated supervision. In practice, there are mainly two methods: (1) Supervised Fine-Tuning (SFT), which teaches the reranker to make ranking decisions across retrieved passages (*e.g.,* relevance scores, orderings), and (2) Reasoning Distillation, which is achieved by training the student to mimic the structured intermediate rationales that the teacher model generates to justify its ranking decisions.

Supervised Fine-Tuning. From a pointwise perspective, LimRank (Song et al., 2025b) generates positive/negative documents derived from long CoT answers to capture implicit relationships between documents and queries. However, ERank (Cai et al., 2025b) argues that binary relevance training leads to poor score discrimination, and replaces it with generative SFT that outputs fine-grained integer scores to better separate subtly different candidates. Beyond traditional SFT, a cold-start SFT teaches reranker an output format, for instance, reasoning patterns (<think> and <answer>) (Liu et al., 2025b) and within-group comparison format (Sun et al., 2025).

Distillation. InteRank (Samarinas and Zamani, 2025) and Reason-to-Rank (Ji et al., 2025) improve reasoning skills by distilling ranking explanations from a teacher-LLM, emphasizing that generating explanations is crucial for effective ranking. Rank1 (Weller et al., 2025c) and Rank-K (Yang et al., 2025) distill reasoning traces into smaller rerankers and enable longer inference-time CoT for the reasoning intensive retrieval queries, yielding stronger performance on BRIGHT. At a finer granularity of reasoning, DeAR (Abdallah et al., 2025) introduces token-level relevance distillation, achieving high accuracy on rerank task.

4.3.3 Reinforcement Learning

Building on SFT and distillation that largely imitate labeled preferences, RL further aligns both *what* the model ranks and *how* it justifies decisions by optimizing task-level rewards tied to ranking quality, output structure, and explanation usefulness. Recent rerankers largely share a Group Relative Policy Optimization (GRPO) backbone, but they diverge in *how* the reward specifies the target behavior, ranging from strict rule checks to richer, metric-driven objectives. At the minimalist end, Rank-R1 (Zhuang et al., 2025) uses a strict rule-check reward on the best-document label, enabling reranker reasoning with only a small amount of reasoning-free labeled data. In contrast, InteRank (Samarinas and Zamani, 2025) automatically generates reward value from a reasoning LLM-based reward model. Beyond single-objective rewards, composite rewards that jointly enforce optimizing ranking from multiple aspects. For instance, REARANK (Zhang et al., 2025a) and TFRank (Fan et al., 2025) combine a score-based reward and a format reward to encourage better-structured, reasoning-centric outputs. To inject broader ranking awareness, ERank (Cai et al., 2025b) and GroupRank (Sun et al., 2025) augment pointwise scoring with listwise-derived rewards computed over the entire candidate list (or groups), encouraging the scorer to respect global ordering. Finally, moving beyond one-shot ranking, ReasonRank (Liu et al., 2025b) optimizes a *multi-view* ranking reward that accounts for the multi-turn nature of sliding-window listwise ranking (combining signals and ranking-similarity measures), so RL explicitly refines end-to-end list quality rather than single-window gains.

4.4 Reasoning-Driven Iterative Retrieval

Reasoning injected into a single retrieval stage can improve performance on reasoning-intensive IR tasks, but naively chaining multiple reasoning modules may amplify redundant "overthinking" and introduce misaligned or drifting reasoning traces. Consequently, *reasoning-driven iterative retrieval* has emerged as a way to coordinate reasoning across stages, refining the search process through adaptive iterations. SMR (Lee et al., 2025a), for example, enforces a state-machine structure that moves from granular token-level analysis to explicit retrieval actions (*e.g.,* REFINE, RERANK, STOP). Similarly, both Li et al. (2025e) and Vijay et al. (2025) cast retrieval as a test-time, iterative decision process guided by an LLM; notably, Vijay et al. (2025) implements this guidance as an RL-trained multi-turn retrieval policy with turn-level rewards and reports stronger effectiveness even with a smaller LLM backbone. In an end-to-end setting, Wang et al. (2025) propose approaches where embedding models iteratively infer and retrieve within the model, progressively sharpening relevance for complex queries, without necessarily retraining the retriever for each refinement step.

5 Open Challenges and Future Directions

Despite rapid progress in RIR, this section examines remaining challenges and future directions.

Evaluation Overly Relies on Traditional IR Metrics. Current evaluation protocols (Su et al., 2025; Li et al., 2025b) still rely primarily on conventional IR metrics such as nDCG and Recall. This introduces two limitations: (1) Efficiency is largely overlooked. Some methods (Long et al., 2025; Chen et al., 2025b) achieve strong effectiveness through complex frameworks but incur high computational costs. Recently, some studies (Peng et al., 2025) have proposed evaluating both the efficiency and effectiveness of current rerankers, while others (Zhou et al., 2024; Weller et al., 2025a; Song et al., 2025a) have introduced metrics tailored to instruction-following retrieval. Moving forward, however, we will likely need novel metrics specifically designed for reasoning-intensive scenarios (*e.g.,* DeepResearch). (2) Fine-grained relevance is not well captured. Two models may obtain similar nDCG scores while retrieving qualitatively different results. Thus, metrics that jointly consider effectiveness and efficiency, as well as fine-grained relevance assessment, are promising directions (Zhang et al., 2025e).

The Domain Generalization Gap in Evaluation. Most RIR benchmarks are anchored in specialized professional settings, from STEM (Su et al., 2025), to legal (Zheng et al., 2025) and medical benchmarks (Li et al., 2025a). Although these resources provide evidence-rich scenarios for structured reasoning (see Table 4), their distance from everyday information needs limits their coverage of broader retrieval tasks. Recent works (Kim et al., 2025; Taghavi et al., 2025) take a step in this direction by testing intent resolution from implicit queries over chat histories, but limited in scale and task diversity. A key next step is scalable, heterogeneous evaluation with broader coverage and stronger generalizability, grounded in routine human–AI interactions.

Bridging the Multimodal Reasoning Gap. Most existing RIR research is confined to text-only (Zhou et al., 2025), whereas integrating visual modalities introduces inferential complexity. Recent multimodal benchmarks (Zhou et al., 2025; Zhang et al., 2026) have extended RIR to vision-language scenarios, revealing a pronounced gap in current MLLMs (Jiang et al., 2024; Zhang et al., 2025d), when tasked with reasoning over joint visual-textual evidence (*e.g.,* spatial relations, causal structure). From a retriever-capability perspective, progress depends on perceptually faithful and fine-grained phrase-to-region grounding, compositional representations that encode explicit cross-model rationales, and the ability to aggregate reasoning across interleaved multi-image evidence.

Inference Latency and Cost. Many high-performing approaches rely on complex multi-stage (Long et al., 2025) or reasoning-enhanced (Chen et al., 2025b) pipelines, resulting in high inference costs. This issue partly stems from the limited reasoning capacity of compact embedding representations and the constraints introduced by contrastive learning. Developing methods (*e.g.,* latent reasoning (Jin et al., 2026) or multi-vector representations (Khattab and Zaharia, 2020)) that balance effectiveness and efficiency would significantly improve practical deployment. More broadly, adaptive routing can allocate reasoning budget based on query difficulty or scenario to control cost without uniformly sacrificing quality.

Generalization Bottlenecks and Narrow Application Scope. Although several works demonstrate cross-benchmark generalization by evaluating on both RIR and traditional IR benchmarks, these specialized methods still often underperform compared to strong general-purpose embedding models (Lee et al., 2025b; Zhang et al., 2025e; Akram et al., 2026). Furthermore, current RIR research mainly focuses on retrieval and reranking. However, RIR naturally aligns with broader applications such as long-term memory systems and deep research assistants. For instance, when a scientist asks a complex research question, a reasoning retriever could leverage historical interests and prior publications to provide personalized and context-aware evidence. Expanding RIR to these practical scenarios presents both evaluation and methodological opportunities.

6 Conclusion

This survey provides a structured roadmap for the rapidly evolving field of RIR. It systematizes the fragmented landscape of benchmarks and datasets, providing a detailed characterization of their difficulty, knowledge domains, and modalities, while introducing a comprehensive reasoning-type taxonomy with examples and an analysis of the reasoning-type focus for each benchmark. We introduce a fine-grained taxonomy that organizes approaches based on where reasoning is incorporated into the retrieval pipeline, spanning pre-retrieval augmentations, retriever training, advanced reranking, and iterative workflows. To contextualize these paradigms, we synthesize theoretical analyses of optimization objectives and provide empirical comparisons for performance and model backbone, mapping these methods to relevant tasks and applications. Finally, we identify key challenges including evaluation metrics innovation, generalization bottlenecks in evaluation and methodologies, bridging the multimodal reasoning gap, and alleviating inference computational costs to make LLM-driven reasoning practical. Addressing these issues is essential for developing the next generation of search systems that are generalizable, reasoning-capable, and practically deployable at scale.

Limitations

While this survey provides an up-to-date and comprehensive review on reasoning-intensive retrieval, we acknowledge several limitations of this survey. First, we mainly include methods that have been empirically evaluated on established reasoning-intensive retrieval benchmarks. Other promising directions (*e.g.,* graph-based retrieval and Hypothetical Document Embeddings, HyDE) are not discussed in depth. Second, our review is restricted to publicly accessible literature and resources, which may overlook proprietary systems and unpublished industrial advances.

References

Abdelrahman Abdallah, Mohamed Darwish Mounis, Mahmoud Abdalla, Mahmoud Salaheldin Kasem, Mostafa Farouk Senussi, Mohamed Mahmoud, Mohammed Ali, Adam Jatowt, and Hyun Soo Kang. 2026. Mm-bright: A multi-task multimodal benchmark for reasoning-intensive retrieval. *ArXiv*, abs/2601.09562.

Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani, and Adam Jatowt. 2025. DeAR: Dual-stage document reranking with reasoning agents via LLM distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5710–5723, Suzhou, China. Association for Computational Linguistics.

Mohammad Kalim Akram, Saba Sturua, Nastia Havriushenko, Quentin Herreros, Michael Günther, Maximilian Werk, and Han Xiao. 2026. jina-embeddings-v5-text: Task-targeted embedding distillation. *ArXiv*, abs/2602.15547.

Shaden Alshammari, Kevin Wen, Abrar Zainal, Mark Hamilton, Navid Safaei, Sultan Albarakati, William T. Freeman, and Antonio Torralba. 2025. Mathnet: a global multimodal benchmark for mathematical reasoning and retrieval. In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*.

Dhananjay Ashok, Suraj Nair, Mutasem Al-Darabsah, Choon Hui Teo, Tarun Agarwal, and Jonathan May. 2025. A representation sharpening framework for zero shot dense retrieval. *arXiv preprint arXiv:2511.05684*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Fengyu Cai, Tong Chen, Xinran Zhao, Sihao Chen, Hongming Zhang, Sherry Tongshuang Wu, Iryna Gurevych, and Heinz Koeppl. 2025a. Revela: Dense retriever learning via language modeling. *ArXiv*, abs/2506.16552.

Yuzheng Cai, Yanzhao Zhang, Dingkun Long, Mingxin Li, Pengjun Xie, and Weiguo Zheng. 2025b. Erank: Fusing supervised fine-tuning and reinforcement learning for effective and efficient text reranking. *arXiv preprint arXiv:2509.00520*.

Hung-Ting Chen, Xiang Liu, Shauli Ravfogel, and Eunsol Choi. 2025a. Beyond single embeddings: Capturing diverse targets with multi-query retrieval. *ArXiv*, abs/2511.02770.

Jianlyu Chen, Junwei Lan, Chaofan Li, Defu Lian, and Zheng Liu. 2025b. Reasonembed: Enhanced text embeddings for reasoning-intensive document retrieval. *arXiv preprint arXiv:2510.08252*.

Liyang Chen, Yujun Cai, Jieqiong Dong, and Yiwei Wang. 2025c. Bright+: Upgrading the bright benchmark with marcus, a multi-agent rag clean-up suite. *arXiv preprint arXiv:2506.07116*.

Peter Baile Chen, Tomer Wolfson, Michael Cafarella, and Dan Roth. 2025d. Enrichindex: Using llms to enrich retrieval indices offline. *arXiv preprint arXiv:2504.03598*.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Jimmy Lin, Akari Asai, and Victor Zhong. 2026. Agentir: Reasoning-aware retrieval for deep research agents. *ArXiv*, abs/2603.04384.

Debrup Das, Sam O'Nuallain, and Razieh Rahimi. 2025. RaDeR: Reasoning-aware dense retrieval models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19981–20008, Suzhou, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix Faltings, Wei Wei, and Yujia Bao. 2025. Enhancing retrieval systems with inference-time logical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 449–463, Vienna, Austria. Association for Computational Linguistics.

Yongqi Fan, Xiaoyang Chen, Dezhi Ye, Jie Liu, Haijin Liang, Jin Ma, Ben He, Yingfei Sun, and Tong Ruan. 2025. Tfrank: Think-free reasoning enables practical pointwise llm ranking. *arXiv preprint arXiv:2508.09539*.

Aniketh Garikaparthi, Manasi Patwardhan, Aditya Sanjiv Kanade, Aman Hassan, Lovekesh Vig, and Arman Cohan. 2025. MIR: Methodology inspiration retrieval for scientific research problems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28614–28659, Vienna, Austria. Association for Computational Linguistics.

Jiahui Geng, Fengyu Cai, Shaobo Cui, Qing Li, Liangwei Chen, Chenyang Lyu, Haonan Li, Derui Zhu, Walter Pretschner, Heinz Koeppl, et al. 2025. Coquir: A comprehensive benchmark for code quality-aware information retrieval. *arXiv preprint arXiv:2506.11066*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nat.*, 645(8081):633–638.

Nilesh Gupta, Wei-Cheng Chang, Ngot Bui, Cho-Jui Hsieh, and Inderjit S Dhillon. 2025. Llm-guided hierarchical retrieval. *arXiv preprint arXiv:2510.13217*.

Jerry Huang, Siddarth Madala, Cheng Niu, J. Hockenmaier, and Tong Zhang. 2025. Contextual relevance and adaptive sampling for llm-based document reranking. *ArXiv*, abs/2511.01208.

Hamel Husain, Hongqiu Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Codesearchnet challenge: Evaluating the state of semantic code search. *ArXiv*, abs/1909.09436.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Yuelyu Ji, Zhuochun Li, Rui Meng, and Daqing He. 2025. Reason-to-rank: Distilling direct and comparative reasoning from large language models for document reranking. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 2320–2329. ACM.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. E5-v: Universal embeddings with multimodal large language models. *ArXiv*, abs/2407.12580.

Jiajie Jin, Yanzhao Zhang, Mingxin Li, Dingkun Long, Pengjun Xie, Yutao Zhu, and Zhicheng Dou. 2026. Laser: Internalizing explicit reasoning into latent space for dense retrieval. *ArXiv*, abs/2603.01425.

Haocheng Ju and Bin Dong. 2025. MIRB: Mathematical information retrieval benchmark. In *2nd AI for Math Workshop @ ICML 2025*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Hyunseo Kim, Sangam Lee, Kwangwook Seo, and Dongha Lee. 2025. BESPOKE: Benchmark for search-augmented large language model personalization via diagnostic feedback. *arXiv preprint arXiv:2509.21106*.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025. Ume-r1: Exploring reasoning-driven generative multimodal embeddings. *arXiv preprint arXiv:2511.00405*.

Dohyeon Lee, Yeonseok Jeong, and Seung-won Hwang. 2025a. From token to action: State machine reasoning to mitigate overthinking in information retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7048–7064, Suzhou, China. Association for Computational Linguistics.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Abrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel M. Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shah Jalalur Rahman Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. 2025b. Gemini embedding: Generalizable embeddings from gemini. *ArXiv*, abs/2503.07891.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. 2024. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

Sangam Lee, Ryang Heo, SeongKu Kang, and Dongha Lee. 2025c. Imagine all the relevance: Scenario-profiled indexing with knowledge expansion for dense retrieval. *arXiv preprint arXiv:2503.23033*.

Yibin Lei, Tao Shen, and Andrew Yates. 2025. ThinkQE: Query expansion via an evolving thinking process. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17772–17781, Suzhou, China. Association for Computational Linguistics.

Lei Li, Xiangxu Zhang, Xiao Zhou, and Zheng Liu. 2025a. AutoMIR: Effective zero-shot medical information retrieval without relevance labels. In *Findings of the Association for Computational Linguistics:*

EMNLP 2025, pages 24028–24047, Suzhou, China. Association for Computational Linguistics.

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

Qingquan Li, Yiran Hu, Feng Yao, Chaojun Xiao, Zhiyuan Liu, Maosong Sun, and Weixing Shen. 2023. Muser: A multi-view similar case retrieval dataset. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 5336–5340, New York, NY, USA. Association for Computing Machinery.

Xiangyang Li, Kuicai Dong, Yi Quan Lee, Wei Xia, Hao Zhang, Xinyi Dai, Yasheng Wang, and Ruiming Tang. 2025c. CoIR: A comprehensive benchmark for code information retrieval models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22074–22091, Vienna, Austria. Association for Computational Linguistics.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025d. From matching to generation: A survey on generative information retrieval. *ACM Trans. Inf. Syst.*, 43(3).

Xingxuan Li, Weiwen Xu, Ruochen Zhao, Fangkai Jiao, Shafiq Joty, and Lidong Bing. 2025e. Can we further elicit reasoning in LLMs? critic-guided planning with retrieval-augmentation for solving challenging tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25589–25604, Vienna, Austria. Association for Computational Linguistics.

Xingxuan Li, Weiwen Xu, Ruochen Zhao, Fangkai Jiao, Shafiq Joty, and Lidong Bing. 2025f. Can we further elicit reasoning in LLMs? critic-guided planning with retrieval-augmentation for solving challenging tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25589–25604, Vienna, Austria. Association for Computational Linguistics.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Chunkit Chan, Yankai Chen, Zhongfen Deng, Yinghui Li, Hai-Tao Zheng, Dongyuan Li, Renhe Jiang, Ming Zhang, Yangqiu Song, and Philip S. Yu. 2025g. A survey of RAG-reasoning systems in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12120–12145, Suzhou, China. Association for Computational Linguistics.

Jintao Liang, Gang Su, Huifeng Lin, You Wu, Rui Zhao, and Ziyue Li. 2025. Reasoning rag via system 1 or system 2: A survey on reasoning agentic retrieval-augmented generation for industry challenges. *ArXiv*, abs/2506.10408.

Junyong Lin, Lu Dai, Ruiqian Han, Yijie Sui, Ruilin Wang, Xingliang Sun, Qinglin Wu, Min Feng, Hao Liu, and Hui Xiong. 2025. Scirgen: Synthesize realistic and large-scale RAG dataset for scientific research. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 5619–5630. ACM.

Yijie Lin, Guofeng Ding, Hao Zhou, Haobin Li, Mouxing Yang, and Xi Peng. 2026. Ark: A dual-axis multimodal retrieval benchmark along reasoning and knowledge. *ArXiv*, abs/2602.09839.

Hongjun Liu, Yilun Zhao, Arman Cohan, and Chen Zhao. 2025a. Sucea: Reasoning-intensive retrieval for adversarial fact-checking through claim decomposition and editing. *ArXiv*, abs/2506.04583.

Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou. 2025b. Reasonrank: Empowering passage ranking with strong reasoning ability. *arXiv preprint arXiv:2508.07050*.

Yuxiang Liu, Tian Wang, Gourab Kundu, Tianyu Cao, Guang Cheng, Zhen Ge, Jianshu Chen, Qingjun Cui, and Trishul Chilimbi. 2025c. Exploring reasoning-infused text embedding with large language models for zero-shot dense retrieval. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025*, pages 4981–4985. ACM.

Meixiu Long, Duolin Sun, Dan Yang, Junjie Wang, Yecheng Luo, Yue Shen, Jian Wang, Hualei Zhou, Chunxiao Guo, Peng Wei, et al. 2025. Diver: A multi-stage approach for reasoning-intensive information retrieval. *arXiv preprint arXiv:2508.07995*.

Shubham Kumar Nigam, Navansh Goel, and Arnab Bhattacharya. 2022. nigam@coliee-22: Legal case retrieval and entailment using cascading of lexical and semantic-based models. In *New Frontiers in Artificial Intelligence: JSAI-IsAI 2022 Workshop, JURISIN 2022, and JSAI 2022 International Session, Kyoto, Japan, June 12–17, 2022, Revised Selected Papers*, page 96–108, Berlin, Heidelberg. Springer-Verlag.

Tong Niu, Shafiq Joty, Ye Liu, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Judgerank: Leveraging large language models for reasoning-intensive reranking. *arXiv preprint arXiv:2411.00142*.

Hanseok Oh, Hyunji Lee, Seonghyeon Ye, Haebin Shin, Hansol Jang, Changwook Jun, and Minjoon Seo. 2024. Instructir: A benchmark for instruction following of information retrieval models. *ArXiv*, abs/2402.14334.

Zhiyuan Peng, Ting-Ruen Wei, Tingyu Song, and Yilun Zhao. 2025. Efficiency-effectiveness reranking FLOPs for LLM-based rerankers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2782–2791, Suzhou (China). Association for Computational Linguistics.

Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, Rui Min, Minpeng Liao, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents. *ArXiv*, abs/2509.13309.

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

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Jon Saad-Falcon, Daniel Y. Fu, Simran Arora, Neel Guha, and Christopher Ré. 2024. Benchmarking and building long-context retrieval models with loco and m2-bert. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Chris Samarinas and Hamed Zamani. 2025. Distillation and refinement of reasoning in small language models for document re-ranking. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, ICTIR 2025, Padua, Italy, 18 July 2025*, pages 430–435. ACM.

Rahul Seetharaman, Kaustubh D Dhole, and Aman Bansal. 2025. Insertrank: Llms can reason over bm25 scores to improve listwise reranking. *arXiv preprint arXiv:2506.14086*.

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. 2025. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*.

Yuchen Shi, Yuzheng Cai, Siqi Cai, Zihan Xu, Lichao Chen, Yulei Qin, Zhijian Zhou, Xiang Fei, Chaofan Qiu, Xiaoyu Tan, Gang Li, Zongyi Li, Haojia Lin, Guocan Cai, Yong Mao, Yunsheng Wu, Ke Li, and Xing Sun. 2025. Youtu-agent: Scaling agent productivity with automated generation and hybrid policy optimization. *ArXiv*, abs/2512.24615.

Tingyu Song, Guo Gan, Mingsheng Shang, and Yilun Zhao. 2025a. IFIR: A comprehensive benchmark for evaluating instruction-following in expert-domain information retrieval. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10186–10204, Albuquerque, New Mexico. Association for Computational Linguistics.

Tingyu Song, Yilun Zhao, Siyue Zhang, Chen Zhao, and Arman Cohan. 2025b. LimRank: Less is more for reasoning-intensive information reranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20636–20650, Suzhou, China. Association for Computational Linguistics.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Liu Haisu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. 2025. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *The Thirteenth International Conference on Learning Representations*.

Duolin Sun, Meixiu Long, Dan Yang, Yihan Jiao, Zhehao Tan, Jie Feng, Junjie Wang, Yue Shen, Peng Wei, Jian Wang, et al. 2025. Grouprank: A groupwise reranking paradigm driven by reinforcement learning. *arXiv preprint arXiv:2511.11653*.

Zeinab Sadat Taghavi, Ali Modarressi, Yunpu Ma, and Hinrich Schuetze. 2025. ImpliRet: Benchmarking the implicit fact retrieval challenge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33156–33178, Suzhou, China. Association for Computational Linguistics.

Jianting Tang, Dongshuai Li, Tao Wen, Fuyu Lv, Dan Ou, and Linli Xu. 2025a. Large reasoning embedding models: Towards next-generation dense retrieval paradigm. *arXiv preprint arXiv:2510.14321*.

Tian Tang, Zhixing Tian, Zhenyu Zhu, Chenyang Wang, Haiqing Hu, Guoyu Tang, Lin Liu, and Sulong Xu. 2025b. Lref: A novel llm-based relevance framework for e-commerce search. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 468–475, New York, NY, USA. Association for Computing Machinery.

Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. 2025. Freshstack: Building realistic benchmarks for evaluating retrieval on technical documents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Supriti Vijay, Aman Priyanshu, Anu Vellore, Baturay Saglam, and Amin Karbasi. 2025. Think before you retrieve: Learning test-time adaptive search with small language models. *arXiv preprint arXiv:2511.07581*.

Guangzhi Wang, Kai Li, Yinghao Jiao, and Zhi Liu. 2025. Refine thought: A test-time inference method for embedding model reasoning. *arXiv preprint arXiv:2511.13726*.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2025a. FollowIR: Evaluating and teaching information retrieval models to follow instructions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the*

Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11926–11942, Albuquerque, New Mexico. Association for Computational Linguistics.

Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. 2025b. Promptriever: Instruction-trained retrievers can be prompted like language models. In *The Thirteenth International Conference on Learning Representations*.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025c. Rank1: Test-time compute for reranking in information retrieval. *arXiv preprint arXiv:2502.18418*.

Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. 2024. Rar-b: Reasoning as retrieval benchmark. *arXiv preprint arXiv:2404.06347*.

Haike Xu and Tong Chen. 2025. Beyond sequential reranking: Reranker-guided search improves reasoning intensive retrieval. *arXiv preprint arXiv:2509.07163*.

Kaishuai Xu, Wenjun Hou, Yi Cheng, and Wenjie Li. 2025. RAR^2: Retrieval-augmented medical reasoning via thought-driven retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20386–20396, Suzhou, China. Association for Computational Linguistics.

Ruiran Yan, Zheng Liu, and Defu Lian. 2025. O1 embedder: Let retrievers think before action. *arXiv preprint arXiv:2502.07555*.

Eugene Yang, Andrew Yates, Kathryn Ricci, Orion Weller, Vivek Chari, Benjamin Van Durme, and Dawn Lawrie. 2025. Rank-k: Test-time reasoning for listwise reranking. *arXiv preprint arXiv:2505.14432*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Yichen Yao, Jiahe Wan, Yuxin Hong, Mengna Zhang, Junhan Yang, Zhouyu Jiang, Qing Xu, Kuan Lu, Yinghui Xu, Wei Chu, and Yuan Qi. 2025. Inf-x-retriever. `https://yaoyichen.github.io/INF-X-Retriever`.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, Online. Association for Computational Linguistics.

Jinsung Yoon, Junhao Zeng, and Sercan O Arik. 2025. SQUARE: Unsupervised retrieval adaptation via synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7283–7297, Suzhou, China. Association for Computational Linguistics.

Hao Yu, Shenyang Huang, Zachary Yang, Maximilian Puelma Touzel, Kellin Pelrine, Jean-François Godbout, and Reihaneh Rabbany. 2025. TRUTH: Teaching LLMs to rerank for truth in misinformation detection. In *Workshop on Socially Responsible Language Modelling Research*.

Le Zhang, Bo Wang, Xipeng Qiu, Siva Reddy, and Aishwarya Agrawal. 2025a. REARANK: Reasoning re-ranking agent via reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2458–2471, Suzhou, China. Association for Computational Linguistics.

Meishan Zhang, Xin Zhang, Xinping Zhao, Shouzheng Huang, Baotian Hu, and Min Zhang. 2025b. On the role of pretrained language models in general-purpose text embeddings: A survey. *ArXiv*, abs/2507.20783.

Siyue Zhang, Yuan Gao, Xiao Zhou, Yilun Zhao, Tingyu Song, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2026. MRMR: A realistic and expert-level multidisciplinary benchmark for reasoning-intensive multimodal retrieval. In *The Fourteenth International Conference on Learning Representations*.

Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2025c. Diffusion vs. autoregressive language models: A text embedding perspective. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4273–4303, Suzhou, China. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025d. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9274–9285.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025e. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *ArXiv*, abs/2506.05176.

Yichi Zhang, Jun Bai, Zhixin Cai, Shuhan Qin, Zhuofan Chen, Jinghua Guan, and Wenge Rong. 2025f. Your dense retriever is secretly an expeditious reasoner. *arXiv preprint arXiv:2510.21727*.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law, CSLAW 2025, Munich, Germany, March 25-27, 2025*, pages 169–193. ACM.

Yunfei Zhong, Jun Yang, Yixing Fan, Lixin Su, Maarten de Rijke, Ruqing Zhang, and Xueqi Cheng. 2025. Reasoning-enhanced query understanding through decomposition and interpretation. *arXiv preprint arXiv:2509.06544*.

Jianqun Zhou, Yuanlei Zheng, Wei Chen, Qi Zheng, Zeyuan Shang, Wei Zhang, Rui Meng, and Xiaoyu Shen. 2024. Beyond content relevance: Evaluating instruction following in retrieval models. *ArXiv*, abs/2410.23841.

Junjie Zhou, Ze Liu, Lei Xiong, Jin-Ge Yao, Yueze Wang, Shitao Xiao, Fenfen Lin, Miguel Hu Chen, Zhicheng Dou, Siqi Bao, et al. 2025. Mr^2-bench: Going beyond matching to reasoning in multimodal retrieval. *arXiv preprint arXiv:2509.26378*.

Changtai Zhu, Siyin Wang, Ruijun Feng, Kai Song, and Xipeng Qiu. 2025. ConvSearch-r1: Enhancing query reformulation for conversational search with reasoning via reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26558–26575, Suzhou, China. Association for Computational Linguistics.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. LongEmbed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 802–816, Miami, Florida, USA. Association for Computational Linguistics.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*.

A Literature Review Procedure

To ensure transparency and rigor, we provide the paper collection strategy and paper selection strategy in this section.

Databases. We searched major sources including ACL Anthology, OpenReview (ICLR, NeurIPS), arXiv, Semantic Scholar, DBLP, and Google Scholar, Github. AI search tools such as PASA, Litmap Connected Papers are also included.

Search Strategy. We applied keyword combinations such as "reasoning retriever," and "retrieval reasoning," within the time range 2024–2026. In addition, we adopt a snowballing strategy by tracing the references and citations of seminal works (*e.g.,* BRIGHT (Su et al., 2025)) and recent contributions (*e.g.,* R2Med (Li et al., 2025b), ReasonIR (Shao et al., 2025)).

Paper Selection Strategy. We have two main principles for selecting suitable papers: (1) The paper must directly address both reasoning and retrieval or search. (2) The paper must be publicly available as a journal article, conference paper, or preprint. Additionally, we will not select papers that: (1) Are abstracts, short articles, or non-academic blog posts. (2) Do not have an accessible full text.

Screening and Statistics. Our initial screening retrieved approximately 400 articles. After deduplication, around 300 articles remained. Applying the inclusion criteria and exclusion yielded 118 papers. After careful human validation of each paper, we finally selected out 56 qualified papers in this domain.

Methodological Rigor. Our protocol is informed by established guidelines for systematic reviews. These emphasize transparent reporting of search strings, last search date, de-duplication, per-stage counts, and inclusion/exclusion flows. By following these standards, we ensure that our literature review process is rigorous, reproducible, and aligned with recognized best practices.

B Reasoning Type Definition

The following reasoning paradigms characterize how domain knowledge in documents supports query resolution. Each type is defined with its logical mechanism and concrete examples.

Deductive Reasoning. A general principle or theorem in the document is directly applied to explain a specific scenario or solve a problem in the query. *Example:* In Table 4 (General/BRIGHT), meristem regeneration theory explains post-cut tree sprouting.

Analogical Reasoning. A document draws a parallel with the query in its underlying logic, indicating that the query and document share a solution strategy or a common theorem/algorithmic foundation. *Example:* In Table 4 (Code/COIR), a C++ Levenshtein distance implementation guides a Python solution via algorithmic equivalence.

Causal Reasoning. The document identifies root causes or mechanistic relationships that explain effects observed in the query. Resolution requires tracing causal chains from symptoms to origins. *Example:* In Table 4 (Code/BRIGHT), missing debug messages are traced to launch file log-level configurations.

Analytical Reasoning. The document provides critical domain knowledge that fills gaps in multi-step reasoning chains required to resolve the query. This involves decomposition of complex problems into interdependent sub-questions. *Example:* In Table 4 (General/BRIGHT), soil science knowledge about salt accumulation completes the reasoning chain for plant water reuse safety.

Numerical Reasoning. The query is resolved by applying quantitative constraints in the document, requiring arithmetic computation (*e.g.,* percentages, unit conversion, rate/ratio) or time arithmetic (*e.g.,* duration, scheduling offsets, temporal comparisons). The logical mechanism is a deterministic mapping from numeric facts and rules to a target value or decision. *Example:* In Table 4 (General/ImpliRet), the document states "Prada Galleria costs $2,000" and "Gucci Marmont is 20% cheaper," so computing $2000 \times 0.8 = 1600$ identifies the Gucci Marmont as the $1,600 bag.

C Complex Retrieval Tasks

To rigorously define the scope of Reasoning-Intensive Retrieval, it is essential to distinguish it from other established complex retrieval paradigms. While these tasks share the need for capabilities beyond simple keyword matching, they differ fundamentally in their core objectives and the nature of the query-document connection.

C.1 Types and Definitions

Multi-Hop Retrieval Multi-hop retrieval addresses scenarios where answering a query requires finding a chain of supporting facts (Yang et al., 2018). The questions are mostly complex that no single document can resolve (*e.g.,* Document A mentions an entity X, and Document B provides the target attribute of X).

Instruction-Following Retrieval Instruction-following retrieval evaluates a retriever's ability to adhere to complex, explicit constraints provided in the user query (Oh et al., 2024). For example, detailed directives regarding length, format, style, or negative constraints (*e.g.,* retrieve documents about apples but exclude any mention of technology).

Long-Context Retrieval Long-context retrieval focuses on the challenge of identifying relevant information ("needles") buried within extremely long inputs ("haystacks"), such as entire books or long legal contracts (Zhu et al., 2024), aiming to test the fidelity of embedding models over extended sequence lengths (*e.g.,* 32k+ tokens) . The core difficulty lies in the *scale* of the context rather than the complexity of the reasoning.

C.2 Comparison with RIR

While complex retrieval tasks involve intricate constraints, they often rely on features explicitly specified in the query (*e.g.,* specific entity attributes in multi-hop retrieval, formatting constraints in instruction-following retrieval). Consequently, these tasks can often be addressed through precise lexical matching (*e.g.,* BM25) or surface-level semantic alignment. In contrast, Reasoning-Intensive Retrieval is defined by relevance signals that are mediated through *latent logical inference chains* (see examples in Table 4). Because the connection is implicit rather than explicitly stated, RIR necessitates a retriever capable of performing reasoning to bridge the gap, rather than relying solely on surface-level overlap.

D Empirical Analysis of RIR Methods.

Beyond categorizing RIR methods, evaluating their practical deployability requires analyzing their empirical performance. This section analyzes the inherent trade-offs between computational overhead, reasoning capacity, and downstream ranking effectiveness. Specifically, we compare the roles of different base models and examine the steep scaling costs associated with multi-stage inference.

LLM-Based vs. LRM-Based Methods. Large Reasoning Models (LRMs) are more suitable for "thinking-heavy" stages, such as complex query rewriting (Guo et al., 2025) and reranking (Liu et al., 2025b), where deeper multi-step inference is required and slightly higher latency is tolerable. In contrast, standard LLMs typically serve as the backbone of the core retrieval stage due to stricter latency constraints, where efficiency and scalability are critical. However, LRMs remain critical offline; they curate high-quality, reasoning-intensive training data to fine-tune standard retrievers to better capture latent logical relevance (Long et al., 2025). Furthermore, some frontier approaches (Zhang et al., 2025f; Jin et al., 2026) have also explored transferring reasoning capabilities from LRMs to LLM architectures through techniques such as distillation, aiming to achieve a better trade-off between effectiveness and efficiency.

Computation Cost vs. Performance. Table 5 and Table 6 summarize the framework and performance among methods. Additionally, we provide computational overhead across different methods in Table 5.

To compare the efficiency of different models, we follow the closed-form formulation of E2R-FLOPs (Peng et al., 2025) and instantiate the cost using each model's architectural hyperparameters, including the number of layers, hidden size, feed-forward dimension, and attention configuration. For single-vector embedding backbones, we estimate the cost of one forward pass using an effective input length defined as the average of the query length and document length, i.e., $(L_q + L_d)/2$, which reflects the mean encoding cost under our corpus statistics. For reranking backbones, we estimate prompt-side FLOPs according to the reranking paradigm: pointwise methods process one query-document pair per call, groupwise and setwise methods process groups of five documents per call, and listwise methods process windows of twenty documents per call. The total computational overhead is then obtained by multiplying the per-call FLOPs by the corresponding number of calls required to rank the top candidates. In this way, our comparison normalizes efficiency across heterogeneous backbones and inference strategies under a unified, hardware-agnostic FLOPs metric.

Foundational single-stage dense retrievers, operating via standard dot-product scoring, deliver robust performance in both effectiveness and robustness. Notably, the strongest retriever, ReasonEmbed-8B (Chen et al., 2025b) even outperforms a $4\times$ larger reranker ReasonRank-32B (Liu et al., 2025b) on both benchmarks.

However, maximizing effectiveness often requires transitioning to multi-stage reasoning architectures. For instance, adding reasoning-aware rerankers yields higher nDCG scores, and multi-step agentic pipelines even achieve peak metrics. On the other hand, they escalate inference costs to the 10^{14} and 10^{16} FLOPs respectively. Thus, frontier approaches seek a middle ground to enhance performance while bounding or reducing compute. In the reranking domain, GroupRank (Sun et al., 2025) combines pointwise computational efficiency with listwise contextual effectiveness. Furthermore, within multi-stage pipelines, INF-X-Retriever (Yao et al., 2025) achieves state-of-the-art performance without compute-heavy rerankers, by directly pairing an intent-recognizing query aligner with a highly optimized retriever.

E Loss Function

InfoNCE (Information Noise-Contrastive Estimation). InfoNCE is a standard objective for self-supervised contrastive learning. Given a query embedding q, a matched positive document d^+, and a set of negatives D^- (optionally including hard negatives), the loss is

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp\big(s(q, d^+)/\tau\big)}{\sum_{d \in \{d^+\} \cup D^-} \exp\big(s(q, d)/\tau\big)} . \quad (1)$$

where $s(\cdot, \cdot)$ denotes a similarity score and $\tau > 0$ is a temperature hyperparameter.

InfoNCE trains retrievers to minimize the representation distance between relevant pairs (*e.g.,* logically related documents in RIR). Curating a high-quality, reasoning-intensive dataset is therefore essential for effective optimization. Specifically, hard negatives are critical for teaching the model to penalize documents that possess surface-level semantic relevance but are logically unrelated to the query (Long et al., 2025; Shao et al., 2025).

Generation Loss. In multi-task training for LLM-based retrievers, a generation objective is commonly used to produce intermediate thoughts or reasoning traces conditioned on the query. For instance i, let q_i be the query/prompt tokens and t_i the target thought tokens; define $x_i = [q_i;\ t_i]$ with $L_i = |q_i| + |t_i|$. The loss typically supervises only the target span:

$$\mathcal{L}_{\text{gen}} = -\sum_{i=1}^{N} \sum_{j=|q_i|+1}^{L_i} \log p_\theta(x_{i,j} \mid x_{i,<j}) , \quad (2)$$

where $x_{i,<j} = (x_{i,1}, \ldots, x_{i,j-1})$.

While standard generation loss optimizes autoregressive next-token prediction, recent LLM-based retrievers repurpose this objective to explicitly train intermediate reasoning steps (Tang et al., 2025a; Lan et al., 2025). Furthermore, LaSER (Jin et al., 2026) advances this approach by internalizing these reasoning patterns directly into the latent embedding space.

Mean Squared Error (MSE). MSE is commonly used for representation matching (*e.g.,* embedding distillation). Given input embeddings $e_i \in \mathbb{R}^d$ and target embeddings $e_i^\star \in \mathbb{R}^d$, a parametric mapping $\mathcal{M}(\cdot; \theta)$ is trained by

$$\mathcal{L}_{\text{MSE}} = \frac{1}{M} \sum_{i=1}^{M} \|\mathcal{M}(e_i; \theta) - e_i^\star\|_2^2 . \quad (3)$$

MSE helps distilling an LLM's deep reasoning capabilities into a computationally cheap embedding space. By training a compact mapper to minimize the distance between a raw query's embedding and its LLM-reasoned counterpart, the system internalizes the semantic transformations of multi-step inference (Zhang et al., 2025f).

F Relevant Tasks and Applications

Reasoning-intensive retrieval extends IR from superficial lexical / semantic relevance to latent inferential link, providing logically grounded evidence for complex tasks. For example, in users' intent recognition task, it improves aligning implicit user query and target corpus by detailed reasoning thought (Chen et al., 2026; Zhu et al., 2025). Additionally, as a part of RAG, it enhances RAG performance by retrieving high-quality documents for truth grounding (Shao et al., 2025). In knowledge-intensive domains, it improves misinformation detection (Yu et al., 2025), fact check (Liu et al., 2025a), scientific paper research (Garikaparthi et al., 2025), complex QA (Liu et al., 2025b), contextual relevance judgment (Ji et al., 2025; Huang et al., 2025), grounding responses in retrieved knowledge and thereby mitigating hallucinations. Reasoning-intensive IR is increasingly applied across diverse domains, including healthcare, software engineering, and e-commerce. The following sections explore domain-specific adaptations of these techniques in greater depth.

Medicine Addressing the complexities of reasoning-intensive retrieval in the medical domain, the RAR^2 (Xu et al., 2025) framework improves diagnostic accuracy by generating an intermediate "thought process" that uncovers implicit clinical knowledge requirements to explicitly guide both the retrieval of evidence and the subsequent reasoning generation.

E-Commerce In the realm of e-commerce reasoning-intensive retrieval, LREM (Tang et al., 2025a) leverages reasoning-then-embedding approach effectively links implicit user queries with intended products, leading to more precise and meaningful retrieval. Additionally, LREF (Tang et al., 2025b) optimizes retrieval performance by utilizing reasoning processes to achieve a more meticulous and granular alignment of query-product relevance.

Software Engineering To address the intricacies of software engineering, reasoning-intensive retrieval improves performance by shifting from static semantic matching to a dynamic process of structural code exploration and verified algorithmic reasoning. CR-Planner (Li et al., 2025f) significantly improves performance on rigorous tasks like competitive programming by employing a critic-guided planning framework to iteratively validate and refine both retrieval queries and reasoning steps, ensuring that generated code is grounded in accurate, verified evidence LATTICE (Gupta et al., 2025) addresses the scalability challenges of searching massive software repositories by imposing a semantic tree structure on the corpus, enabling the LLM to actively traverse hierarchical paths and efficiently pinpoint deeply nested logic that flat retrieval methods often miss.

Domain	Name	Size	Data Source	Query	Doc	Reflecting Real-World Difficulty
Open Domain	ImpliRet (Taghavi et al., 2025)	9,000	Internet, LLM	Natural Text	Chat history	Document-side reasoning with no lexical overlap
	BESPOKE (Kim et al., 2025)	150	Human	Natural Text	Chat history	Capture implicit user preferences in multi-turn chat APP
Scientific	MIRB (Ju and Dong, 2025)	39,029	Internet, Math Libraries, Previous Dataset	Natural/Formal Text	Theorem, Formula, Proof, Question	Automated math theorem proving
	MathNet-Retrieve (Alshammari et al., 2025)	10,000	Contest, Human, LLM	Formal Text/Image	Similar Question	Retrieve mathematically equivalent problems in multilingual and multimodal domains.
	ScIRGen (Lin et al., 2025)	61,376	Internet, LLM, Papers	Natural Text	Paper Content	Complex task-oriented research questions in scientific workflows
	FreshStack (Thakur et al., 2025)	672	Internet, LLM	Natural Text	Document, Code	Find realistic solutions from niche, up-to-date technical documents
Code	CoIR (Li et al., 2025c)	≈162,000	Contest, Human, Internet, LLM, Previous Dataset	Natural Text, Code	Code, Answer	Code Summary, Code Translation
	CoQuIR (Geng et al., 2025)	42,725	Internet, LLM, Previous Dataset	Natural text	Code	Prioritizing quality over mere functional relevance
Legal	Legal-Benchmark (Zheng et al., 2025)	9,863	Databases, Human, Internet, Textbooks	Natural Text	Answer, Statute	Quick search for relevant statutes based on realistic legal issues.
Medical	R2MED (Li et al., 2025b)	876	Human, Internet, LLM, Papers, Previous Dataset, Textbooks	Natural Text	Answer, Document, Diagnosis	Explore complete latent diagnoses and treatment planning from symptoms for doctors
	CMIRB (Li et al., 2025a)	10,962	Internet, LLM, Papers, Previous Dataset	Natural Text	Document, Diagnosis, Question	Match patient symptoms to consultations
Multi-Domain	BRIGHT (Su et al., 2025)	1,384	Contest, LLM, Human, Internet, Previous Dataset, Textbooks	Natural Text	Theorem, Code, Document, Question	Find supportive evidence with deeper logical connection (*e.g.,* scientific search)
	BRIGHT+ (Chen et al., 2025c)	1,384	Contest, LLM, Human, Internet, Previous Dataset, Textbooks	Natural Text	Theorem, Code, Document, Question	(same as BRIGHT)
	RAR-b (Xiao et al., 2024)	45,745	Internet, Previous Dataset	Natural Text	Answer, Code	Automated answer annotation on scientific QA
Multi-Modal	MRMR (Zhang et al., 2026)	1,435	LLM, Human, Internet, Previous Dataset	Natural Text/Image	Answer, Theorem, Document, Image	Expert-level visual interpretation and interleaved modalities
	MR2-BENCH (Zhou et al., 2025)	1,309	LLM, Human, Internet, Papers, Previous Dataset, Textbooks	Natural Text/Image	Document, Theorem, Diagram, Image	Understand and retrieve content in complex and multi-modal document structure.
	ARK (Lin et al., 2026)	1,547	LLM, Human, Internet, Previous Dataset, Papers	Natural Text/Image	Image, Diagram, Chart, Scientific Illustration	Abstract conceptual connection between knowledge and scientific documents.
	MM-BRIGHT (Abdallah et al., 2026)	2,803	LLM, Human, Internet	Natural Text/Image	Document, Image, Multimodal	Reasoning-intensive multi-task retrieval from real expert technical queries with integral images.

Table 3: Overview of existing reasoning-intensive retrieval benchmarks.

Domain	Benchmark	Reasoning Type	Example	Inference Chain
Open Domain	ImpliRet	Numerical	**Query:** "Which bag costs \$1,600?" **Related Doc:** "The Prada Galleria costs \$2,000; the Gucci Marmont is 20% cheaper."	Given reference price (\$2,000) $\rightarrow$ Apply discount rule (20% cheaper $\Rightarrow 0.8 \times 2000 = 1600$) $\rightarrow$ Match target amount (\$1,600 $\Rightarrow$ Gucci)
Scientific	MIRB	Deductive (Symbolic)	**Query:** "Open covering H of closed bounded S in R has finite subcover He from H" **Related Doc:** No point in S^c is limit point of S	Theorem (Heine-Borel for compactness) $\rightarrow$ Prerequisite (S closed and bounded) $\rightarrow$ Property (closed: no exterior limit points)
	MathNet-Retrieve	Analytical (Multimodal)	**Query:** "Prove points D, E, F, G, H are concyclic..." **Related Doc:** Proof by drawing EG and FH, chasing equal angles...	Core problem $\rightarrow$ Analysis root (parallelogram) $\rightarrow$ Implications (parallel lines) $\rightarrow$ Conclusion (equal angles force concyclicity)
	BRIGHT (Biology)	Deductive	**Query:** "After cutting trees into logs... they grow normal stems..." **Related Doc:** Document on meristematic tissues	Phenomenon $\rightarrow$ Supportive theory (cell division) $\rightarrow$ Applied concept (meristem)
Code	BRIGHT (Robotics)	Causal	**Query:** "Can't see debug messages using RCLCPP_DEBUG..." **Related Doc:** Launch file with log_level default 'info'...	Symptom $\rightarrow$ Potential cause (node log level override) $\rightarrow$ Configuration (default 'info' arg)
	COIR	Analogical	**Query:** Python code implementing Levenshtein distance...	Query pattern $\rightarrow$ Algorithmic equivalence $\rightarrow$ Language translation $\rightarrow$ Structural mapping
Legal	LegalBench	Deductive	**Query:** "Teacher fired from private school..." **Related Doc:** 14th Amendment Due Process...	Legal issue $\rightarrow$ Supportive Rule $\rightarrow$ Rules application $\rightarrow$ Facts connection $\rightarrow$ Conclusion
Medical	R2MED	Analytical	**Query:** "An 82-year-old woman... What is the next test?" **Related Doc:** Video-capsule endoscopy...	Core problem $\rightarrow$ Analysis root $\rightarrow$ Latent reasoning $\rightarrow$ Diagnostic method
	CMIRB	Deductive	**Query:** "How long after thyroid surgery can one return to work?" **Related Doc:** Healing timeline...	Phenomenon $\rightarrow$ Supportive healing process $\rightarrow$ Relevant timeline
Multimodal	MRMR	Deductive	**Query:** "Jack was driving through..." **Image:** A white car crossing lane... **Related Doc:** "Driving in tunnels — Rule (f)"	Observed behavior $\rightarrow$ Applicable regulation $\rightarrow$ Constraint violation $\rightarrow$ Relevant document retrieval
	MRMR	Causal	**Query:** "What causes black bulges on a corn cob?"	Visual symptom $\rightarrow$ Potential cause $\rightarrow$ Specific cause $\rightarrow$ Disease identification

Table 4: Examples of domain-specific benchmarks with key reasoning types, query examples, and inference chains.

Role	Method	Backbone	Size	Framework	Inference	nDCG@10	FLOPs
Group 1: Single-Stage Retrieval							
Retriever	ReasonEmbed (Chen et al., 2025b)	Qwen-3	8B	Single Vector	Dot Product	**38.1**	2.0584e12
			4B	Single Vector	Dot Product	37.1	1.1001e12
	DIVER-Retriever (Long et al., 2025)	Qwen-3	4B	Single Vector	Dot Product	28.9	1.1001e12
	RaDeR (Das et al., 2025)	GTE-Qwen2	7B	Single Vector	Dot Product	25.5	1.8511e12
	ReasonIR (Shao et al., 2025)	Llama-3.1	8B	Single Vector	Dot Product	24.4	2.0443e12
Group 2: Reranking & Multi-Stage Pipelines							
Reranker	GroupRank (Sun et al., 2025)	Qwen-2.5	32B	DIVER-4B + GPT-4 query rewrite	CoT Generation	**39.2**	2.2687e15
			7B	DIVER-4B + GPT-4 query rewrite	CoT Generation	36.7	4.7405e14
	ReasonRank (Zhang et al., 2025a)	Qwen-2.5	32B	ReasonIR-8B + GPT-4 query rewrite	CoT Generation	38.0	2.2207e15
			7B	ReasonIR-8B + GPT-4 query rewrite	CoT Generation	35.7	4.6486e14
	Rank-R1 (Zhuang et al., 2025)	Qwen-2.5	7B	with BM25	CoT Generation	16.4	4.7405e14
	TFRank (Fan et al., 2025)	Qwen-3	1.7B	with BM25	Think-Free	16.7	1.4556e14
			0.6B	with BM25	Think-Free	15.6	5.1896e13
Pipeline	INF-X-Retriever (Yao et al., 2025)	GTE-Qwen2	7B	Query Aligner + Retriever	Multi-Step	**63.4**	-
	DIVER (Long et al., 2025)	Qwen-3	8B	Expander + Retriever + Reranker	Multi-Step	46.8	-

Table 5: Representative reasoning-intensive retrieval methods and their performance on the **BRIGHT** benchmark. Best score in each subgroup is in **bold**.

Role	Key Method	Backbone	Size	Framework Design	nDCG@10
Group 1: Retrieval *(Single-Stage Dense Retrieval)*					
Retriever	ReasonEmbed (Chen et al., 2025b)	Qwen-3	8B	Single Vector	**43.18**
			4B	Single Vector	41.16
	DIVER-Retriever* (Long et al., 2025)	Qwen-3	4B	Single Vector	42.91
Group 2: Reranking *(Multi-Stage)*					
Reranker	GroupRank (Sun et al., 2025)	Qwen-2.5	32B	with DIVER-Retriever-4B	**52.28**
			7B	with DIVER-Retriever-4B	47.84
	ReasonRank (Zhang et al., 2025a)	Qwen-2.5	32B	with E5-mistral-7B	42.85
			7B	with E5-mistral-7B	39.53

* DIVER-Retriever data is from GroupRank paper.

Table 6: Representative Reasoning-Intensive Retrieval Methods Overview and Performance Landscape on **R2MED** benchmark. The top performance in each subgroup is highlighted in **bold**.