

# Rethinking Reasoning-Intensive Retrieval: Evaluating and Advancing Retrievers in Agentic Search Systems

Yilun Zhao<sup>Y</sup>

Jinbiao Wei<sup>Y</sup>

Tingyu Song<sup>Y</sup>

Siyue Zhang<sup>S</sup>

Chen Zhao<sup>N</sup>

Arman Cohan<sup>Y</sup>

<sup>Y</sup>Yale NLP Lab

<sup>S</sup>National University of Singapore

<sup>N</sup>NYU Shanghai

 RTriever

 BRIGHT-PRO

## Abstract

Reasoning-intensive retrieval aims to surface evidence that supports downstream reasoning rather than merely matching topical similarity. This capability is increasingly important for agentic search systems, where retrievers must provide complementary evidence across iterative search and synthesis. However, existing work remains limited on both evaluation and training: benchmarks such as BRIGHT provide narrow gold sets and evaluate retrievers in isolation, while synthetic training corpora often optimize single-passage relevance rather than evidence portfolio construction. We introduce BRIGHT-PRO, an expert-annotated benchmark that expands each query with multi-aspect gold evidence and evaluates retrievers under both static and agentic search protocols. We further construct RTriever-Synth, an aspect-decomposed synthetic corpus that generates complementary positives and positive-conditioned hard negatives, and use it to LoRA fine-tune RTriever-4B from Qwen3-Embedding-4B. Experiments across lexical, general-purpose, and reasoning-intensive retrievers show that aspect-aware and agentic evaluation expose behaviors hidden by standard metrics, while RTriever-4B substantially improves over its base model.

## 1 Introduction

Information Retrieval (IR) has long served as the foundation for accessing and organizing knowledge at scale (Singh et al., 2025; Ajith et al., 2024; Kamaloo et al., 2024). Traditional IR systems have achieved remarkable success in scenarios where user intent can be satisfied by retrieving factoid or single-hop information (Zhang et al., 2025d; Weller et al., 2024; Lee et al., 2024). However, as user queries become increasingly complex and demand multi-step reasoning, retrieval systems must move beyond surface-level relevance, giving rise

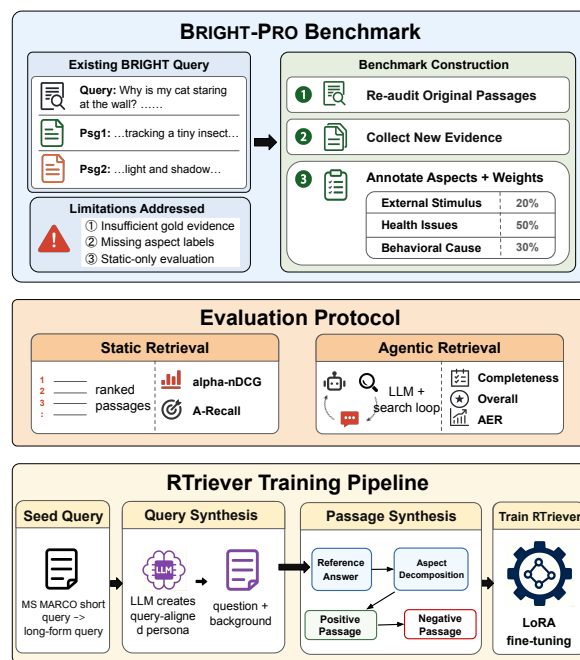


Figure 1: Overview of our work. *Top*: BRIGHT-PRO augments BRIGHT with re-audited gold passages and aspect-level labels, addressing gaps in evidence coverage and evaluation. *Middle*: retrievers are evaluated under both static and agentic protocols. *Bottom*: RTriever is trained via LoRA on synthesized queries and aspect-decomposed positive/negative passages.

to *reasoning-intensive retrieval*, a new frontier in IR research (Xiao et al., 2024; SU et al., 2025).

In practice, reasoning-intensive queries often require multi-step reasoning and the integration of diverse evidence, which causes retrievers to struggle to surface sufficiently relevant information within a single retrieval step. To address this limitation, recent work has turned to complex, multi-stage information-seeking pipelines, culminating in the emergence of *Deep-Research* (Chen et al., 2025b; Du et al., 2025; Yang et al., 2025b; Wu et al., 2025; Zhao et al., 2026). Deep-research systems employ LLM-based agents that iteratively plan, search, read, and synthesize information in

order to answer difficult queries, largely compensating for the limited effectiveness of integrated retrievers. Because relevant evidence is rarely surfaced in a single retrieval step, these systems rely on increased iterations of retrieval and reasoning, leading to substantial computational cost and latency. This reliance highlights the importance of reasoning-intensive retrieval: a retriever that surfaces high-quality, reasoning-ready evidence in a single step would reduce iteration cost and improve deep-research efficiency.

Despite growing interest in reasoning-intensive retrieval, both evaluation and training remain misaligned with this agentic use case. On the evaluation side, the main existing benchmark, BRIGHT (SU et al., 2025), provides each query with only a small set of gold passages, typically derived from one or two web pages, and evaluates retrievers in isolation rather than inside a dynamic deep-research loop. On the training side, recent synthetic corpora for reasoning retrieval (Shao et al., 2025b; Long et al., 2025) often pair each query with a single positive and hard negatives, encouraging models to rank *a* relevant passage highly rather than to retrieve complementary evidence that jointly supports complex reasoning. As a result, a retriever can appear strong under single-passage metrics while still failing to cover the reasoning aspects needed by an agent.

To bridge the evaluation gap, we introduce **BRIGHT-PRO**, an evaluation framework that extends BRIGHT with richer, multi-aspect supervision for reasoning-intensive retrieval. Expert annotators expand each query’s gold passage set and group passages into *reasoning aspects* that reflect distinct perspectives or subproblems within the same query, as illustrated in Figure 1. This enables fine-grained analysis of whether a retriever covers the full reasoning need rather than concentrating on a single aspect. Beyond static evaluation, we integrate retrievers into an LLM-based agentic workflow that iteratively plans, retrieves, and synthesizes information, allowing us to measure system-level outcomes such as reasoning completeness, iteration efficiency, and final response quality.

To address the training gap, we construct **RTriever-Synth**, an aspect-decomposed synthetic corpus designed to teach retrievers complementary evidence selection. Starting from MS MARCO seeds, our pipeline rewrites short search queries into realistic analytical queries, generates a self-contained reference answer, decomposes that an-

swer into non-overlapping reasoning aspects, and realizes each aspect as a positive passage. It then generates positive-conditioned hard negatives: passages that share topical cues with the query but deliberately omit the aspect needed to satisfy the information need. We use this corpus to LoRA fine-tune **RTriever-4B** from Qwen3-Embedding-4B (Zhang et al., 2025c).

We evaluate classical lexical, general-purpose, and reasoning-intensive retrievers on BRIGHT-PRO. Across subsets and settings, BGE-Reasoner (Chen et al., 2025a) consistently delivers the strongest performance, while the DIVER family (Long et al., 2025) forms a strong second tier and Qwen-family embedders (Li et al., 2023a; Zhang et al., 2025c) trail by a noticeable gap. Our RTriever-4B lifts the 4B base by a wide margin and lands in the upper-middle tier across settings, ahead of much larger general-purpose embedders. Although BM25 performs poorly in static retrieval, it becomes competitive under the agentic protocol, where targeted follow-up queries reduce vocabulary mismatch and allow lexical matching to surface useful evidence. Case studies further show that early access to gold evidence shortens interaction traces and improves reasoning completeness, whereas weaker retrievers invite speculative reasoning and recycle low-utility passages. Reasoning-intensive retrieval thus calls for joint progress on evaluation and training to support iterative research workflows.

Our main contributions are summarized below:

- We introduce BRIGHT-PRO, an expert-annotated benchmark that extends BRIGHT with multi-aspect evidence and evaluates retrievers under both static and agentic search settings.
- We propose RTriever-Synth, an aspect-decomposed synthesis pipeline that generates complementary positives from reference-answer reasoning and positive-conditioned hard negatives for training retrievers.
- We train RTriever-4B by LoRA fine-tuning Qwen3-Embedding-4B on RTriever-Synth, providing a 4B retriever specialized for reasoning-intensive evidence selection.
- We deliver an empirical study that disentangles retriever quality from agent-retriever fit, showing that aspect-aware and agentic protocols expose retrieval behaviors hidden by standard rankings.

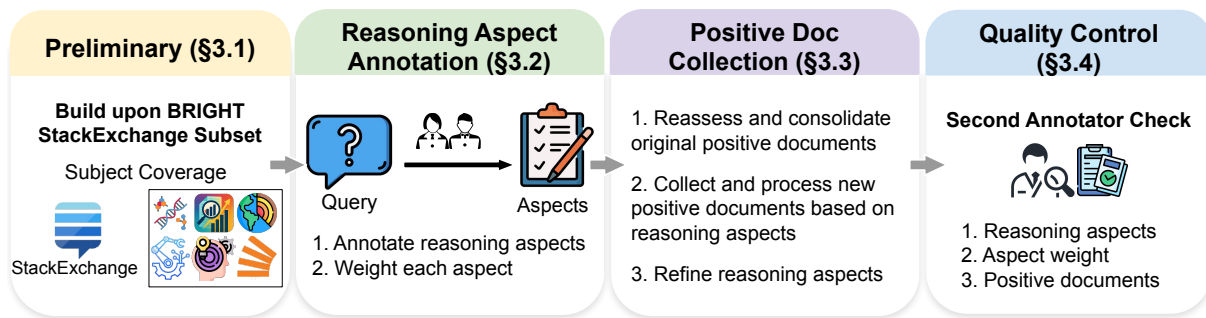


Figure 2: An overview of the BRIGHT-PRO benchmark construction pipeline.

## 2 Related Work

### Reasoning-Intensive Information Retrieval.

Beyond traditional keyword- and semantic-based IR, BRIGHT (SU et al., 2025) is the first benchmark that explicitly targets queries requiring multi-step reasoning to identify genuinely useful evidence rather than merely superficial relevance, with concurrent extensions to multimodal (Abdallah et al., 2026b; Zhang et al., 2025a) and instruction-following (Song et al., 2025a) settings. Building on this formulation, recent work has trained retrievers on synthetic data containing reasoning-intensive queries and hard negatives to encourage reasoning-aware evidence selection (Weller et al., 2025; Das et al., 2025; Zhang et al., 2025b; Long et al., 2025; Shao et al., 2025b; Jin et al., 2026; Chen et al., 2026; Song et al., 2025b). However, current evaluations typically score a retriever against a small pool of gold passages tied to one or two sources (Abdallah et al., 2026a), which limits assessment of coverage across complementary reasoning aspects.

**Agentic Search System and Evaluation.** In parallel, agentic search systems (e.g., DeepResearch) combine LLM planning with iterative search, reading, and synthesis to tackle complex queries (Chen et al., 2025b; Yang et al., 2025b; Wu et al., 2025; Shao et al., 2025a; Ding et al., 2026). Several benchmarks are proposed to evaluate the DeepResearch system responses (Wang et al., 2025; Du et al., 2025; Yifei et al., 2025; Xiong et al., 2026; Li et al., 2026; Gupta et al., 2026; Wu et al., 2026; Hu et al., 2026). To enable controlled comparisons of agent components, BROWSECOMP-PLUS (Chen et al., 2025b) provides a fixed, curated corpus that standardizes access and reduces environmental variance during evaluation. While this design improves reproducibility, it abstracts away open-domain retrieval dynamics and offers limited visibility into

how a retriever shapes an agent’s evidence portfolio, iteration budget, and final reasoning quality, gaps echoed by recent calls for explicit information-coverage evaluation (Samuel et al., 2026).

## 3 BRIGHT-PRO Benchmark

This section introduces BRIGHT-PRO, which builds upon BRIGHT (SU et al., 2025) to enable a more thorough evaluation of reasoning-intensive retrieval. Figure 2 outlines the construction pipeline.

### 3.1 Preliminary: BRIGHT Benchmark

The BRIGHT benchmark is divided into three subsets: (1) *StackExchange*, which focuses on retrieving passages that support answering questions sourced from StackExchange; (2) *Coding*, which focuses on retrieving documentation or analogous solved problems; and (3) *Theorem*, which targets retrieving solved problems that use the same theorems or contain relevant theorem statements. BRIGHT-PRO builds upon the StackExchange subset because it best represents open-domain, natural language reasoning. In contrast, the coding and theorem subsets rely heavily on domain-specific syntax or formal logic, making them less suitable for evaluating general-purpose reasoning retrievers.

For queries in the StackExchange subset, annotators trace hyperlinks in accepted or high-quality answers to collect the specific web pages that those answers depend on. Each web page is then segmented into multiple passages, which serve as gold passages. To construct negative samples, annotators use Google Search with either the post title or LLM-generated keywords to locate topically related web pages that do not meet the precise informational need of the query. All collected passages (i.e., both positive and negative across all queries) are then compiled into a unified retrieval corpus.

### 3.2 Reasoning Aspect Annotation

In real-world reasoning-intensive retrieval scenarios, users seek a comprehensive set of evidence that collectively supports the entire reasoning chain behind a query. Reasoning completeness often decomposes into multiple reasoning aspects, where each aspect represents a coherent subproblem or perspective. When a retriever exhibits bias toward passages aligned with only one aspect, the resulting evidence portfolio becomes unbalanced, thereby weakening downstream synthesis.

**Reasoning Aspect Annotation.** Motivated by this, for each query from the BRIGHT StackExchange subset we assign field-specific expert annotators and ask them to decompose the information need into a compact set of *reasoning aspects*. We provide links to community answers on StackExchange as a starting point. Because user-contributed answers can be concise, audience-tailored, heuristic, or outdated, annotators must think beyond them and independently identify the premises needed for a defensible solution, each documented with a one- to two-sentence rationale.

**Aspect Weight Annotation.** To reflect unequal contributions across aspects, we explicitly weight aspects rather than treating them as interchangeable. This design focuses evaluation on what most affects correctness, reduces gaming by accumulating low-value snippets, and aligns scoring with the utility of the final synthesis. For each reasoning aspect, annotators assign Likert scores from 1 to 5 based on the aspect’s importance to producing a correct and defensible final answer (the detailed rubric is provided in Appendix A). We then normalize the scores to weights  $w_a \in [0, 1]$  with  $\sum_a w_a = 1$ .

### 3.3 Aspect-Guided Positive Doc Collection

After establishing the reasoning aspects and their weights, we proceed to collect positive documents that comprehensively support each query.

**Reassessment and Consolidation of Original Positive Documents.** The BRIGHT benchmark provides a set of positive passages. However, these passages are not assigned to reasoning aspects and, in our preliminary screening, a nontrivial portion was weakly related to the underlying queries. Therefore, we first ask annotators to re-audit the original BRIGHT positives: each passage is reviewed for topical fidelity to the query, and either assigned to current aspects or discarded if

it fails to provide explicit, verifiable evidence. Because many BRIGHT positives were split from the same source page, we further reconcile passages that are contiguous or substantively overlapping. When multiple passages from a single URL support the same aspect, annotators merge them into one coherent segment to preserve context and reduce redundancy, while trimming unrelated text. Each merged segment is stored as a single positive tied to its aspect.

### Acquisition and Processing of New Positive Documents.

Annotators are allowed to use both conventional web search or AI-assisted search like Perplexity AI and ChatGPT with Web Search when searching for aspect-relevant documents. A document is accepted as a positive instance only if it provides explicit, precise, and credible evidence supporting one or more aspects. We provide annotators with a customized interface built upon the FireCrawl framework. Annotators input the URL of a web page, and the interface automatically downloads the HTML source, removes boilerplate content (*e.g.*, advertisements, navigation menus), and extracts the textual content for further review. Annotators are then required to manually refine the extracted text to remove residual noise or irrelevant segments. If a document’s content covers multiple reasoning aspects, annotators must segment the text into aspect-specific portions so that each segment aligns with a single aspect. Each segmented portion is then stored as an independent positive sample associated with its respective aspect.

### Iterative Refinement of Reasoning Aspects.

Document collection is an iterative process that naturally leads to refinement of the aspect structure. As annotators search, they often uncover missing assumptions, dependencies between aspects, or cases where two aspects capture overlapping reasoning needs. In such cases, annotators carefully adjust the aspect schema, for example, by clarifying definitions to eliminate redundancy, subdividing overly broad aspects into more precise ones, or consolidating highly correlated aspects that contribute to the same line of reasoning.

### 3.4 Data Quality Control

Each example in BRIGHT-PRO is independently re-examined by a second annotator from the same field, who verifies both the reasoning aspects (their granularity, balance, weights, and coverage of the reasoning process) and the supporting documents

Subset	Total Number				Avg. Length			Example
	Q	$\mathcal{D}$	$\mathcal{D}^+$	$\mathcal{A}$	Q	$\mathcal{D}$	$\mathcal{A}$	
Biology	103	59,513	7.81	3.94	92.6	75.7	20.1	Tab. 6
Earth Science	115	123,575	7.44	3.83	82.2	41.8	20.1	Tab. 7
Economics	99	52,240	7.81	3.71	123.5	80.4	19.9	Tab. 8
Psychology	100	54,741	7.07	3.84	116.2	72.7	19.8	Tab. 9
Robotics	101	63,920	6.17	3.71	218.8	45.6	19.1	Tab. 10
Stack Overflow	115	109,188	4.60	3.32	172.0	151.8	19.0	Tab. 11
Sustain. Living	106	63,142	9.25	3.86	116.9	69.2	20.2	Tab. 12
Overall	739	526,319	7.13	3.74	131.4	79.3	19.7	—

Table 1: Data statistics of BRIGHT-PRO. For each subset, we show the number of queries ( $\mathbf{Q}$ ) and documents ( $\mathbf{D}$ ), the average number of positive documents per query ( $\mathbf{D}^+$ ) and the average number of reasoning aspects per query ( $\mathbf{A}$ ), together with the average length (in whitespace-separated words) of queries, documents, and aspect descriptions.

(whether each offers explicit, credible evidence for its assigned aspect). To assess weighting reliability, we randomly sample 50 queries and ask independent reviewers to rescore the annotated reasoning aspects; the resulting weighted Cohen’s  $\kappa$  of 0.742 indicates stable importance ratings across annotators. The full review protocol is described in Appendix B. Table 1 presents the key statistics of the BRIGHT-PRO benchmark. Examples for each subset can be found in Appendix G.

## 4 BRIGHT-PRO Evaluation Protocol

We evaluate retrievers along two complementary dimensions: a *static* setting, which isolates retrieval quality using aspect-annotated gold passages, and an *agentic* setting, which assesses the retriever inside a deep-research workflow.

### 4.1 Static Retrieval Evaluation

We first evaluate retrievers in a traditional IR setting, where each query receives a single ranked list of passages. Because every query is annotated with reasoning aspects of varying importance (Section 3.2), we adopt a metric that rewards covering complementary aspects rather than over-retrieving from a single one. Our primary metric is  $\alpha$ -**nDCG@k** (Clarke et al., 2008) with novelty penalty  $\alpha = 0.5$ , which discounts repeated retrieval from already-covered aspects. We complement it with **Weighted Aspect Recall (A-Recall@k)**, which credits each aspect once covered, and report **NDCG@k** and **Recall@k** as diagnostics. See Appendix C for full definitions.

### 4.2 Agentic Pipeline and Response Evaluation

To measure retrievers as components of deep-research systems, we plug each retriever into the same LLM agent equipped with a single search tool. The retriever is the only experimental variable: the agent receives the original query, iteratively issues search queries over the corresponding corpus, reads the returned passages, and produces a citation-grounded final answer. We run this protocol with two agent backends, GPT-5-mini-08-07 and Qwen3.5-122B-A10B (Yang et al., 2025a), using identical prompts, tool interfaces, retrieval depth, and stopping budgets; implementation details are provided in Appendix F, and prompt templates in Appendix E.

We evaluate final answers with GPT-5 as an LLM-as-Judge, following prior work on deep-research evaluation (Chen et al., 2025b; Li et al., 2024). The judge compares each system answer against a reference answer generated from the human-annotated reasoning aspects and their positive passages. It assigns each aspect a coverage score in  $\{0, 0.5, 1\}$  and an answer-level overall quality score from 1 to 5. We aggregate aspect coverage using the annotated aspect weights and report it as reasoning completeness, together with overall quality. Reference construction and validation are detailed in Appendix D; the judge prompt appears in Figure 6.

### 4.3 Fixed- and Adaptive-Round Protocols

To bound cost while preserving comparability, all agentic experiments use the same fixed sample of 175 queries, with 25 queries per task.

**Fixed-Round Protocol.** The agent performs exactly  $R \in \{1, 2, 3\}$  search rounds. Each round returns the top-5 passages, and after each round the agent generates an answer conditioned on all evidence retrieved so far. This protocol isolates retriever quality under matched interaction budgets. We report cumulative  $\alpha$ -**nDCG@k** at  $5R$ , with reasoning completeness and overall answer quality.

**Adaptive-Round Protocol.** The agent decides when to stop searching based on whether the accumulated evidence is sufficient. This protocol measures both answer quality and interaction efficiency: a stronger retriever should satisfy the agent’s evolving information need in fewer rounds. We report the number of retrieval rounds, reasoning completeness, overall quality, and the **Efficiency-Quality**

**Reward (AER):**

$$\text{AER} = OQ \times e^{-\gamma(R-1)} \quad (1)$$

where  $OQ$  is the overall quality of the response,  $R$  is the number of search rounds, and  $\gamma = 0.05$ .

## 5 RTriever: Learning to Retrieve Complementary Evidence

Reasoning-intensive retrieval is rarely consumed in isolation. Inside a deep-research agent, the retriever is queried repeatedly across rounds, and the value of any single retrieval is judged by whether the accumulated set of passages *collectively* satisfies the user’s reasoning need. Yet current synthetic corpora for reasoning-intensive retrieval (Shao et al., 2025b; Long et al., 2025) pair every query with a single synthetic positive, so a contrastive retriever trained on them learns to rank *a* relevant document highest, not to surface a balanced portfolio of complementary evidence. We close this gap with **RTriever-Synth**, a synthetic corpus whose every analytical query is paired with a set of complementary positives that together span the reasoning aspects of the query.

### 5.1 Synthesis Pipeline

The pipeline runs in two stages: it first rewrites each MS MARCO seed into a realistic deep-research query, then decomposes a reference answer into reasoning aspects, realizes each aspect as a complementary positive passage, and synthesizes matched hard negatives.

**Realistic Query Surface from Real Seeds.** MS MARCO queries provide broad topical coverage, but many are short search-engine queries rather than the long-form information needs faced by reasoning-intensive retrievers. To convert these terse seeds into more natural query surfaces while preserving topical control, we start from MS MARCO (Campos et al., 2016) queries as semantic seeds and, for each seed, sample three personas from PersonaHub (Chan et al., 2024) as format examples. We then prompt an LLM to create a new persona aligned with the seed query, and to rewrite the seed query in that persona’s voice as a DeepResearch-style post with a question and short background; the concatenation is what the retriever consumes. A subsequent LLM classifier labels the query as *factual* or *analytical*: factual queries ask for a specific piece of information that can typically be supported by a single passage, whereas

analytical queries require multiple complementary aspects to jointly support the answer, and this label drives branching in the next stage.

**Aspect-Decomposed Gold via Reference Answers.** The central design choice in our pipeline is that we do *not* synthesize positives directly from the query. For each analytical query, we first prompt a strong LLM to produce a comprehensive, self-contained answer of the kind a careful expert would give once all evidence is in hand. We then ask a second LLM call to decompose the reasoning of that answer into two to three *non-overlapping* aspects, each tied to an explicit rationale that names which part of the answer the aspect supplies. For each aspect, the LLM produces a passage blueprint consisting of the rationale, passage type, source, title, and a three-sentence TL;DR. A separate passage-realization call then instantiates each blueprint as a full positive passage in the indicated source style. Two properties follow by construction: every positive is load-bearing for the reasoning chain, and the positives within a query are complementary rather than restatements, so any single passage answers part of the query while the full set jointly answers it. The resulting bundles train the retriever to favor complementary evidence portfolios over isolated high-scoring passages. Factual queries skip decomposition and produce a single positive blueprint.

**Positive-Conditioned Hard Negatives.** After the positive passage blueprints are fixed, we synthesize an equal number of hard negatives conditioned on the query and the positive passage summaries. The negative generator sees the titles and TL;DRs of the positives and is instructed to propose passages that share topical or lexical cues with the query but fail the user’s actual information need when compared with those positives. For each negative, it outputs a rationale, a failure justification, source/type/title metadata, a TL;DR, and an explicit missing aspect to avoid. In the final passage-realization step, positives and negatives are generated from their blueprints using the same source- and genre-conditioned writer; for negatives, the missing aspect is added as an avoidance constraint, yielding passages that remain plausible and topically close while deliberately omitting the evidence needed to answer the query.

Model	BRIGHT Overall	Biology	Earth Science	Economics	Psychology	Robotics	Stack Overflow	Sustainable Living	Overall
BGE-Reasoner-8B	33.8	<b>73.5</b>	<b>74.6</b>	<b>66.0</b>	<b>65.5</b>	<b>64.7</b>	<b>67.7</b>	<b>63.8</b>	<b>68.0</b>
DIVER-4B-1020	30.6	<u>72.8</u>	<u>72.1</u>	61.9	<u>59.9</u>	<u>56.7</u>	<u>63.8</u>	<u>58.8</u>	<u>63.7</u>
DIVER-4B	28.9	67.3	71.2	53.4	<u>60.4</u>	53.1	60.3	53.7	59.9
RTriever-4B (ours)	27.7	63.1	64.7	53.6	50.1	49.2	52.7	53.9	55.3
INF-Retriever-Pro (7B)	26.3	62.6	64.3	50.7	51.9	46.1	50.1	50.6	53.8
Qwen3-8B	23.7	52.7	58.6	46.0	46.6	48.4	52.8	41.0	49.5
INSTRUCTOR-XL (1.5B)	18.9	45.1	54.6	44.2	45.9	44.5	47.5	42.3	46.3
OpenAI-Embed-3L	17.9	53.5	55.2	44.7	44.6	37.6	44.2	40.8	45.8
GTE-7B	22.5	60.3	62.5	39.1	37.4	41.8	40.0	37.5	45.5
GritLM (7B)	21.0	50.6	58.3	40.3	37.6	43.8	38.8	39.2	44.1
ReasonIR-8B	24.4	42.6	52.2	34.9	35.6	40.9	45.7	35.1	41.0
BM25	14.5	41.9	49.1	40.2	30.6	39.2	40.2	40.7	40.3
EmbeddingGemma-300M	18.9	48.7	55.8	32.6	32.7	35.3	35.2	35.7	39.4

Table 2: Performance of retrievers under BRIGHT-PRO static retrieval setting. We report  $\alpha$ -nDCG@25 scores across all subsets, along with the NDCG@10 score for the BRIGHT results. Models are ranked by their overall performance. The best and second-best models for each subset are shown in bold and underlined, respectively.

## 5.2 RTriever Training Details

From one million MS MARCO queries we randomly sample 140K, and the synthesis pipeline generates 140K complete (query, positives, negatives) bundles after filtering. For each training step, we pair every query with one randomly-sampled positive passage and one randomly-sampled negative passage.

We obtain **RTriever-4B** by LoRA fine-tuning Qwen3-Embedding-4B (Zhang et al., 2025c) on this sampled subset of RTriever-Synth, using the MS-SWIFT (Zhao et al., 2024b) framework. LoRA adapters are attached to all linear projection layers of the backbone with rank  $r=16$  and scaling factor  $\alpha=32$ , while the original embedding parameters remain frozen. Training optimizes a contrastive INFONCE objective with temperature  $\tau=0.02$  on query–document pairs: each query is paired with its positive document and one synthesized hard negative, and other documents in the same batch additionally serve as in-batch negatives. We do not include query–query or document–document contrastive pairs. The model is trained for 5 epochs with a peak learning rate of  $1 \times 10^{-5}$ , a 5% linear warm-up, and bf16 mixed-precision optimization through DeepSpeed ZeRO-2; we use a per-device batch size of 384 over 2 NVIDIA B200 GPUs (effective batch 768) with no gradient accumulation, FlashAttention-2, and padding-free packing. Sequences are truncated to 2,048 tokens.

## 6 Experiment

We organize our experiments around two questions: how do existing retrievers perform on reasoning-intensive queries under static and agentic evalua-

tion, and what failure patterns emerge inside an agentic search loop?

### 6.1 Experiment Setup

We benchmark BRIGHT-PRO against twelve publicly available retrievers in addition to our RTriever-4B, organized into three categories: (1) *Classical Lexical Baselines*: BM25 (Robertson et al., 2009). (2) *General-Purpose Retrievers*: GritLM (7B) (Muennighoff et al., 2024), INSTRUCTOR-XL (1.5B) (Su et al., 2022), GTE-Qwen2-7B-Instruct (GTE-7B) (Li et al., 2023b), Qwen3-Embedding-8B (Zhang et al., 2025c), EmbeddingGemma-300M (Vera et al., 2025), and OpenAI text-embedding-3-Large (Neelakantan et al., 2022). (3) *Reasoning-Intensive Retrievers*: ReasonIR-8B (Shao et al., 2025b), DIVER-Retriever-4B together with its more recent checkpoint DIVER-Retriever-4B-1020 (Long et al., 2025), BGE-Reasoner-Embed-Qwen3-8B (Chen et al., 2025a), and INF-Retriever-v1-Pro (7B) (Yang et al., 2025c). For each baseline retriever, we employ its native tokenizer, embedding interface, and pooling strategy without any further fine-tuning. Input sequences are truncated to each model’s validated maximum context length. For instruction-tuned retrievers, we apply either the same prompts used in BRIGHT or model-recommended instructions where applicable to ensure fair and consistent evaluation. All evaluation experiments are conducted on a cluster of NVIDIA H200 GPUs.

### 6.2 Main Findings

We organize the main findings around three evaluation settings: static retrieval, fixed-round agentic retrieval, and adaptive-round agentic retrieval.

Model	Round 1			Round 2			Round 3		
	$\alpha$ -nDCG (0-100)	Compl. (1-5)	Overall (1-5)	$\alpha$ -nDCG (0-100)	Compl. (1-5)	Overall (1-5)	$\alpha$ -nDCG (0-100)	Compl. (1-5)	Overall (1-5)
BGE-Reasoner-8B	<b>56.66</b>	<b>4.14</b>	<b>3.95</b>	<b>61.35</b>	<b>4.40</b>	<b>4.26</b>	<b>63.04</b>	<b>4.42</b>	<b>4.31</b>
DIVER-4B	46.75	3.95	3.78	<u>50.29</u>	4.21	4.12	53.08	4.38	4.29
RTriever-4B (ours)	44.22	<u>3.98</u>	<u>3.85</u>	48.17	4.17	4.12	50.79	4.37	4.25
GTE-7B	45.88	3.93	3.84	49.80	4.17	<u>4.13</u>	52.68	4.33	4.23
DIVER-4B-1020	44.28	3.90	3.77	49.21	4.22	4.06	51.56	4.33	4.16
INSTRUCTOR-XL	35.24	3.54	3.50	40.15	4.09	3.97	43.48	4.26	4.14
BM25	44.46	3.82	3.67	48.05	4.19	4.05	51.48	4.25	4.12
ReasonIR-8B	37.14	3.81	3.65	41.28	4.06	3.93	44.90	4.16	4.11
Qwen3-8B	44.11	3.81	3.74	48.77	<u>4.25</u>	4.11	51.75	4.26	4.10
GritLM (7B)	40.28	3.89	3.74	43.99	4.19	4.03	47.05	4.24	4.07

Table 3: Performance of retrievers under BRIGHT-PRO fixed-round agentic evaluation setting (GPT-5-mini agent). Because each round returns top-5 passages, we report  $\alpha$ -nDCG at the cumulative cut-off:  $k=5$  after round 1,  $k=10$  after round 2, and  $k=15$  after round 3. **Compl.**: reasoning completeness of the final response; **Overall**: overall quality of the final response. Rows are ordered by round-3 Overall quality (descending).

**Static Retrieval.** Aspect-aware evaluation creates a clear separation that BRIGHT NDCG@10 (adjacent column in Table 2) does not: four reasoning-intensive retrievers (BGE-Reasoner-8B, DIVER-4B-1020, DIVER-4B, INF-Retriever-Pro) plus our RTriever-4B form a tight upper tier 4–14 points above every general-purpose embedder, including 8B-parameter Qwen3-Embedding-8B and OpenAI text-embedding-3-Large. Training objective dominates parameter count: RTriever-4B (4B) beats every 7–8B general-purpose retriever, while ReasonIR-8B (although reasoning-trained) falls to eleventh of thirteen at  $\alpha$ -nDCG@25 = 41.0 because its single-positive-per-document pipeline optimizes the BRIGHT-style single-passage signal rather than multi-aspect coverage.

**Fixed-Round Agentic Retrieval.** Static rankings translate only loosely to the agentic loop. BGE-Reasoner-8B leads on both axes (Round-3  $\alpha$ -nDCG@15 = 63.0, +9.9 over the next retriever; Overall = 4.31), and DIVER-4B finishes ahead of its newer DIVER-4B-1020 sibling (Overall 4.29 vs. 4.16) despite trailing statically. Our RTriever-4B places third on Overall (4.25), close to DIVER-4B and ahead of every general-purpose retriever. Below the top tier, retrieval rank ( $\alpha$ -nDCG) and answer rank (Overall) diverge; the case studies in §6.3 trace the mechanisms behind this divergence. BM25 climbs from worst statically within this evaluation set (40.3) to a Round-3  $\alpha$ -nDCG@15 of 51.5, and LLM-issued follow-ups with concrete keywords close its vocabulary-mismatch gap, suggesting that part of what “reasoning-intensive retrieval” demands can be off-loaded to the agent, a deployment-relevant pattern that static evaluations fail to surface.

Model	Agent	#R	Compl.	Overall	AER
BGE-Reasoner-8B	GPT-5-mini	<b>5.10</b>	<b>4.63</b>	4.43	<b>3.65</b>
	Qwen3.5	<b>4.14</b>	<u>4.22</u>	3.99	<b>3.44</b>
RTriever-4B (ours)	GPT-5-mini	6.01	4.53	4.43	3.51
	Qwen3.5	<u>4.89</u>	<b>4.26</b>	<b>4.06</b>	<u>3.38</u>
BM25	GPT-5-mini	5.73	4.50	4.42	3.53
	Qwen3.5	5.21	4.19	4.01	3.31
DIVER-4B	GPT-5-mini	5.91	4.57	4.46	3.53
	Qwen3.5	5.56	4.20	<u>4.02</u>	3.29
DIVER-4B-1020	GPT-5-mini	<u>5.61</u>	4.54	4.43	<u>3.56</u>
	Qwen3.5	6.57	4.18	3.96	3.11
Qwen3-8B	GPT-5-mini	6.27	4.52	<u>4.49</u>	3.50
	Qwen3.5	6.16	4.15	3.93	3.14
GTE-7B	GPT-5-mini	6.67	<u>4.62</u>	<b>4.51</b>	3.44
	Qwen3.5	5.83	4.13	3.95	3.18
GritLM (7B)	GPT-5-mini	7.03	4.51	4.47	3.36
	Qwen3.5	6.48	4.08	3.85	3.08
ReasonIR-8B	GPT-5-mini	7.17	4.48	4.42	3.31
	Qwen3.5	6.99	4.10	3.93	3.08
INSTRUCTOR-XL	GPT-5-mini	7.14	4.46	4.35	3.26
	Qwen3.5	6.93	4.05	3.88	3.04

Table 4: Performance of retrievers under BRIGHT-PRO adaptive-round agentic evaluation setting, with two agent backbones: GPT-5-mini and Qwen3.5-122B-A10B. Rows are ordered by the mean AER across the two agents (descending).

**Adaptive-Round Agentic Retrieval.** The adaptive setting separates retrievers that converge quickly from those merely *eventually* good. BGE-Reasoner-8B again leads (AER = 3.65) with the fewest rounds (5.10), and our RTriever-4B is second by mean AER across the two agents (GPT-5-mini AER = 3.51 over 6.01 rounds); in contrast, GTE-7B has the highest overall quality (4.51) yet its 6.67 rounds drop AER to 3.44, the failure mode AER is designed to surface, since deployment cost scales with rounds, not just final answer quality. Switching to the Qwen3.5-122B-A10B agent pre-

serves the top tier but collapses DIVER-4B-1020 from second to seventh (AER 3.56  $\rightarrow$  3.11), while our RTriever-4B *rises* to second (3.38, 0.06 behind BGE-Reasoner). Top-tier rankings remain stable, but lower-tier rankings reflect retriever-agent compatibility more than retriever quality alone.

### 6.3 Qualitative Analysis

Inspecting 175 RTriever-4B + GPT-5-mini adaptive-round traces surfaces five recurring patterns; we describe each below and illustrate it with a real run from the same retriever-agent stack.

**Early-Round Retrieval Efficiency.** When retrieval surfaces most of the gold evidence in the first one or two rounds, the agent can compose a directly grounded answer and terminates without further exploration. For example, Figure 7 shows an Antarctic ice-sheet question resolved in three rounds, with all four reasoning aspects covered.

**Evidence Deprivation and Speculative Reasoning.** When retrieval fails to surface any gold passage, the LLM still has to answer, so it improvises a plausible-sounding response from whatever the retriever returned, and the answer is invariably speculative. We see this in the Gazebo Garden case (Figure 8), which runs thirteen rounds without retrieving either gold document and incorrectly concludes the plugin is deprecated.

**Repetition Bias and Self-Reinforcing Retrieval Errors.** The retriever can lock onto a topic-adjacent cluster in early rounds; subsequent rounds then keep resurfacing the same passages even when the agent rephrases the query in markedly different ways, so the evidence frontier fails to expand. Consider Figure 9, where twelve rounds yield only 28 unique documents from 60 retrieval slots, and the two highest-weighted aspects are never reached.

**Aspect Tunnel Vision.** On multi-aspect questions, every search round can elaborate the same aspect while other required aspects receive zero retrieval coverage. The search queries themselves remain novel each round, so the lock-in is semantic rather than document-level, but the effect on aspect-aware metrics is the same. A representative run appears in Figure 10: all seven rounds drill into one half of a two-part GHCN climate question and never query the other half.

**Hypothesis Hopping after Early Success.** The converse of early termination occurs when the

agent retrieves the gold passages in the first one or two rounds but cannot stop, and instead spends additional rounds testing alternative concept labels that add little to the final answer. As one illustration, in Figure 11 round 2 already retrieves the *Baader–Meinhof phenomenon* gold, yet the run continues for four more rounds, paying the deployment cost that AER (Table 4) penalises.

## 7 Conclusion

We introduce BRIGHT-PRO, a reasoning-intensive retrieval benchmark with multi-aspect gold sets and agentic retriever-in-the-loop protocols, together with RTriever-Synth, an aspect-decomposed synthetic corpus for training retrievers. Our experiments show that aspect-aware evaluation exposes retrieval behaviors hidden by single-passage metrics, and that static retrieval quality does not always predict utility inside agentic search loops. On the training side, RTriever-4B substantially improves over its Qwen3-Embedding-4B base, suggesting that complementary positives and positive-conditioned negatives provide useful supervision for reasoning-intensive retrieval. These findings call for retrievers built and evaluated around full evidence portfolios rather than individual passages.

### Limitations and Future Work

BRIGHT-PRO builds upon the StackExchange subset of BRIGHT, which currently covers seven expert domains. However, this scope may not fully capture the diversity and complexity of real-world, reasoning-intensive retrieval scenarios. Future research could extend our work by incorporating a broader range of expert domains (Zhao et al., 2024a; Shen et al., 2026) to enhance representativeness. Additionally, the number of examples in BRIGHT-PRO remains limited due to the high cost of human expert annotation. Future work could explore semi-automated or hybrid annotation pipelines to scale up data creation while maintaining quality and reliability. Finally, our use of RTriever-Synth is intentionally simple: we fine-tune RTriever-4B on a subset of one-positive-one-negative triplets. The full training corpus enables further exploration of multi-positive objectives (Wang et al., 2026; Esfandiarpour et al., 2025), aspect-aware sampling (Dou et al., 2025), negative curricula (Zhou et al., 2025; Moreira et al., 2025), and how synthetic evidence structure affects reasoning-intensive retrievers (Feng et al., 2026).

## References

- Abdelrahman Abdallah, J. Holdcroft, M. Ali, and Adam Jatowt. 2026a. Are llm-based retrievers worth their cost? an empirical study of efficiency, robustness, and reasoning overhead.
- Abdelrahman Abdallah, Mohamed Darwish Mounis, Mahmoud Abdalla, M. Kasem, Mostafa Farouk Senussi, Mohamed Mahmoud, Mohammed Ali, Adam Jatowt, and H. Kang. 2026b. Mm-bright: A multi-task multimodal benchmark for reasoning-intensive retrieval. *ArXiv*, abs/2601.09562.
- Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. [Lit-search: A retrieval benchmark for scientific literature search](#).
- Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, L. Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.
- Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *ArXiv*, abs/2406.20094.
- Jianlyu Chen, Junwei Lan, Chaofan Li, Defu Lian, and Zheng Liu. 2025a. Reasonembed: Enhanced text embeddings for reasoning-intensive document retrieval. *arXiv preprint arXiv:2510.08252*.
- Zijian Chen, Xueguang Ma, Shengyao Zhuang, Jimmy Lin, Akari Asai, and Victor Zhong. 2026. Agentir: Reasoning-aware retrieval for deep research agents.
- Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, et al. 2025b. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent. *arXiv preprint arXiv:2508.06600*.
- Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. [Novelty and diversity in information retrieval evaluation](#). In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08*, page 659–666, New York, NY, USA. Association for Computing Machinery.
- Debrup Das, Sam O’ Nuallain, and Razieh Rahimi. 2025. [Rader: Reasoning-aware dense retrieval models](#).
- Hang Ding, Yilun Zhao, Tiansheng Hu, Manasi Patwardhan, and Arman Cohan. 2026. Scirag: Adaptive, citation-aware, and outline-guided retrieval and synthesis for scientific literature. pages 6440–6460.
- Zheng Dou, Deqing Wang, Fuzhen Zhuang, Jian Ren, and Yanling Hu. 2025. Flew: Facet-level and adaptive weighted representation learning of scientific documents. *ArXiv*, abs/2509.07531.
- Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.
- Reza Esfandiarpour, George Zerveas, Ruochen Zhang, Macton Mgonzo, Carsten Eickhoff, and Stephen H. Bach. 2025. Beyond contrastive learning: Synthetic data enables list-wise training with multiple levels of relevance. pages 22860–22882.
- Xincan Feng, Noriki Nishida, Yusuke Sakai, and Yuji Matsumoto. 2026. The wisdom of many queries: Complexity-diversity principle for dense retriever training. *ArXiv*, abs/2602.09448.
- Nikita Gupta, Riju Chatterjee, Lukas Haas, Connie Tao, Andrew Wang, Chang Liu, Hidekazu Oiwa, E. Gribovskaya, Jan Ackermann, John Blitzer, S. Goldshtein, and D. Das. 2026. Deepsearchqa: Bridging the comprehensiveness gap for deep research agents. *ArXiv*, abs/2601.20975.
- Tiansheng Hu, Yilun Zhao, Canyu Zhang, Arman Cohan, and Chen Zhao. 2026. Sage: Benchmarking and improving retrieval for deep research agents. *ArXiv*, abs/2602.05975.
- 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.
- Ehsan Kamaloo, Nandan Thakur, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, and Jimmy Lin. 2024. [Resources for brewing beer: Reproducible reference models and statistical analyses](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1431–1440.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. [Llms-as-judges: A comprehensive survey on llm-based evaluation methods](#).
- Ruizhe Li, Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2026. Deepresearch bench ii: Diagnosing deep research agents via rubrics from expert report. *ArXiv*, abs/2601.08536.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023a. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

- Meixiu Long, Duolin Sun, Dan Yang, Junjie Wang, Yue Shen, Jian Wang, Peng Wei, Jinjie Gu, and Jiahai Wang. 2025. [Diver: A multi-stage approach for reasoning-intensive information retrieval](#).
- G. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2025. Improving text embedding models with positive-aware hard-negative mining. *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*.
- Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*.
- Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. [Text and code embeddings by contrastive pre-training](#).
- Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- Saron Samuel, Andrew Yates, Dawn J. Lawrie, Ian Soboroff, Trevor Adriaanse, Benjamin Van Durme, and Eugene Yang. 2026. Coveragebench: Evaluating information coverage across tasks and domains.
- Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. 2025a. [Dr tulou: Reinforcement learning with evolving rubrics for deep research](#).
- 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, and Luke Zettlemoyer. 2025b. [ReasonIR: Training retrievers for reasoning tasks](#). In *Second Conference on Language Modeling*.
- Yifei Shen, Yilun Zhao, Justice Ou, Tinglin Huang, and Arman Cohan. 2026. [Patient-similarity cohort reasoning in clinical text-to-SQL](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1367–1412, Rabat, Morocco. Association for Computational Linguistics.
- Amanpreet Singh, Joseph Chee Chang, Chloe Anastasiades, Dany Haddad, Aakanksha Naik, Amber Tanaka, Angele Zamarron, Cecile Nguyen, Jena D. Hwang, Jason Dunkleberger, Matt Latzke, Smita Rao, Jaron Lochner, Rob Evans, Rodney Kinney, Daniel S. Weld, Doug Downey, and Sergey Feldman. 2025. [Ai2 scholar qa: Organized literature synthesis with attribution](#).
- Tingyu Song, Guo Gan, Mingsheng Shang, and Yilun Zhao. 2025a. Ifir: A comprehensive benchmark for evaluating instruction-following in expert-domain information retrieval. pages 10186–10204.
- Tingyu Song, Yilun Zhao, Siyue Zhang, Chen Zhao, and Arman Cohan. 2025b. Limrank: Less is more for reasoning-intensive information reranking. *ArXiv*, abs/2510.23544.
- Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.
- 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 Thirtieth International Conference on Learning Representations*.
- Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel M. Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jing Zheng, Jyotinder Singh, Abheesht Sharma, Divya Sreepat, Aashi Jain, Adham Elarabawy, A. Co., Andreas Dourmanglou, Babak Samari, B. Hora, B. Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Abrego, He Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, P. Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariaifar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, S. Qiu, T. Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gael Liu, Jiageng Zhang, Kathleen Kenealy, M. Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, C. Brick, Olivier Lacombe, A. Roberts, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. 2025. Embeddingemma: Powerful and lightweight text representations. *ArXiv*, abs/2509.20354.
- Benben Wang, Minghao Tang, Hengran Zhang, Jiafeng Guo, and Keping Bi. 2026. Training dense retrievers with multiple positive passages. *ArXiv*, abs/2602.12727.

- Jiayu Wang, Yifei Ming, Riya Dulepet, Qinglin Chen, Austin Xu, Zixuan Ke, Frederic Sala, Aws Albarghouti, Caiming Xiong, and Shafiq Joty. 2025. [Livere-searchbench: A live benchmark for user-centric deep research in the wild.](#)
- Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025. [Rank1: Test-time compute for reranking in information retrieval.](#)
- Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. 2024. [Promptriever: Instruction-trained retrievers can be prompted like language models.](#) *arXiv preprint arXiv:2409.11136.*
- Bin Wu, Arastun Mammadli, Xiaoyu Zhang, and Emine Yilmaz. 2026. [Agentsearchbench: A benchmark for ai agent search in the wild.](#)
- Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, et al. 2025. [Webwalker: Benchmarking llms in web traversal.](#) *arXiv preprint arXiv:2501.07572.*
- Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. 2024. [Rar-b: Reasoning as retrieval benchmark.](#) *arXiv preprint arXiv:2404.06347.*
- Lei Xiong, Kun Luo, Ziyi Xia, Wenbo Zhang, Jin-Ge Yao, Zheng Liu, Jing Shao, Jianlyu Chen, Hongjin Qian, Xi Yang, Qian Yu, Hao Li, C. Yue, Xia'an Du, Yuyang Wang, Yesheng Liu, Haiyu Xu, and Zhicheng Dou. 2026. [Autoresearchbench: Benchmarking ai agents on complex scientific literature discovery.](#)
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. [Qwen3 technical report.](#)
- John Yang, Kilian Lieret, Carlos E. Jimenez, Alexander Wettig, Kabir Khandpur, Yanzhe Zhang, Binyuan Hui, Ofir Press, Ludwig Schmidt, and Diyi Yang. 2025b. [Swe-smith: Scaling data for software engineering agents.](#)
- Junhan Yang, Jiahe Wan, Yichen Yao, Wei Chu, Yinghui Xu, Emma Wang, and Yuan Qi. 2025c. [inf-retriever-v1 \(revision 5f469d7\).](#)
- Li S. Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. 2025. [Researchqa: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics.](#)
- Siyue Zhang, Yuan Gao, Xiao Zhou, Yilun Zhao, Tingyu Song, Arman Cohan, A. Luu, and Chen Zhao. 2025a. [Mrmr: A realistic and expert-level multidisciplinary benchmark for reasoning-intensive multimodal retrieval.](#) *ArXiv, abs/2510.09510.*
- Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2025b. [Diffusion vs. autoregressive language models: A text embedding perspective.](#) *arXiv preprint arXiv:2505.15045.*
- 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. 2025c. [Qwen3 embedding: Advancing text embedding and reranking through foundation models.](#) *arXiv preprint arXiv:2506.05176.*
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025d. [Qwen3 embedding: Advancing text embedding and reranking through foundation models.](#) *arXiv preprint arXiv:2506.05176.*
- Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. 2024a. [FinDVer: Explainable claim verification over long and hybrid-content financial documents.](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14739–14752, Miami, Florida, USA. Association for Computational Linguistics.
- Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Yixin Liu, Xiangru Tang, Joseph Chee Chang, Jesse Dodge, Jonathan Bragg, Chen Zhao, Hannaneh Hajishirzi, Doug Downey, and Arman Cohan. 2026. [Sciarena: An open evaluation platform for non-verifiable scientific literature-grounded tasks.](#) In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track.*
- Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024b. [Swift: a scalable lightweight infrastructure for fine-tuning.](#) *ArXiv, abs/2408.05517.*
- Jiawei Zhou, Hang Ding, and Haiyun Jiang. 2025. [Ark: Answer-centric retriever tuning via kg-augmented curriculum learning.](#) *ArXiv, abs/2511.16326.*

## A Aspect Importance Rubric

To reflect unequal contributions across reasoning aspects, annotators assign each aspect a Likert-scale score from 1 to 5 based on its importance to producing a correct and defensible final answer. The scores are then normalized into weights  $w_a \in [0, 1]$  with  $\sum_a w_a = 1$ . Table 5 details the grading rubric used during annotation.

## B Data Quality Control Protocol

This section expands on the second-annotator review summarized in the main text.

**Aspect Review.** For each example, reviewers carefully assess the reasoning aspects to verify that they are comprehensive, balanced, and conceptually sound. The aspects should neither be overly granular nor too general, and collectively they should capture the complete reasoning process behind the question. Reviewers also verify that the assigned weight accurately represents each aspect’s contribution to the overall reasoning process, adjusting weights when initial annotations overemphasize minor details or undervalue critical components, to maintain both consistency and conceptual completeness.

**Positive Document Review.** After verifying the reasoning aspects, the reviewers re-examine each positive document to confirm that it provides explicit, credible, and directly relevant evidence for its assigned aspect. Reviewers remove redundant or loosely related content and ensure that every passage offers concrete reasoning support rather than surface-level associations.

## C Static Retrieval Metrics: Notation and Formulas

We provide the formal definitions of the four metrics used in the static retrieval evaluation. As discussed in Section 3.2, each query  $q$  is annotated with  $m$  reasoning aspects  $\{a_1, \dots, a_m\}$ , with each aspect assigned a nonnegative weight  $w_j$  representing its importance to the overall reasoning process. Let  $\mathcal{G}$  be the set of gold passages and let  $a : \mathcal{G} \rightarrow \{1, \dots, m\}$  map each gold passage to its unique aspect. Given a ranked list  $\langle d_1, \dots, d_k \rangle$ ,  $r \in \{1, \dots, k\}$  denotes the rank position and  $d_r$  is the document at rank  $r$ . We define  $C_j(k) = \sum_{r=1}^k \text{rel}_{rj}$  as the count of retrieved gold items for aspect  $j$  within the top  $k$ , and the binary

relevance  $\text{rel}_r = \mathbb{1}\{d_r \in \mathcal{G}\}$ , so that the aspect-specific relevance becomes

$$\begin{aligned} \text{rel}_{rj} &= \mathbb{1}\{d_r \in \mathcal{G} \wedge a(d_r) = j\} \\ &= \text{rel}_r \mathbb{1}\{a(d_r) = j\}. \end{aligned} \quad (2)$$

**$\alpha$ -nDCG@k.**  $\alpha$ -nDCG@ $k$  (Clarke et al., 2008) explicitly models aspect-level diversity, rewarding retrievers that cover multiple reasoning aspects rather than retrieving redundant passages from the same subtopic. The gain at rank  $r$  is

$$G(r) = \sum_{j=1}^m w_j \text{rel}_{rj} (1 - \alpha)^{C_j(r-1)}, \quad (3)$$

where  $\alpha \in [0, 1]$  controls the novelty penalty (we set  $\alpha = 0.5$  throughout the paper). Accumulating with a logarithmic discount,

$$\text{DCG}_\alpha@k = \sum_{r=1}^k \frac{G(r)}{\log_2(r+1)}. \quad (4)$$

For normalization, we compute  $\text{IDCG}_\alpha@k$  on the gold pool using the same gain definition and a greedy maximization over  $k$  positions, yielding

$$\alpha\text{-nDCG}@k = \frac{\text{DCG}_\alpha@k}{\text{IDCG}_\alpha@k}. \quad (5)$$

**Weighted Aspect Recall.** To directly capture aspect coverage, we report a weighted aspect recall that credits each aspect once it has been covered at least once:

$$\text{A-Recall}@k = \sum_{j=1}^m w_j \cdot \mathbb{1}\{C_j(k) \geq 1\}. \quad (6)$$

**Recall@k and NDCG@k.** As complementary metrics that ignore aspect structure, we additionally report the standard Recall@ $k$  (fraction of gold passages within the top  $k$ ) and NDCG@ $k$  (with binary relevance  $\text{rel}_r$ ).

## D Reference Answer Validation

We use GPT-5 with a high reasoning effort setting to generate one citation-grounded reference answer per query. The model is given the human-annotated reasoning aspects together with the full content of the positive passages linked to each aspect, and is prompted to produce a comprehensive answer that cites the supporting passages.

To verify that the synthesized reference answers are reliable enough to serve as the gold target for

Score	Interpretation
5	<b>Dominant:</b> This aspect captures the core reasoning required for a correct answer. It directly determines whether the response’s main claim or logic is valid.
4	<b>Strong:</b> The aspect provides key supporting reasoning that significantly improves correctness or interpretive depth, though not the single decisive factor.
3	<b>Moderate:</b> The aspect is necessary to develop a coherent or complete explanation, but its omission would only moderately weaken correctness.
2	<b>Supporting:</b> The aspect refines or contextualizes the reasoning, helping link ideas or strengthen justification but is secondary in determining correctness.
1	<b>Peripheral:</b> The aspect has minimal influence compared with others but still contributes meaningfully to understanding or completeness.

Table 5: Rubric for assigning weights (1–5) among meaningful reasoning aspects.

LLM-as-Judge scoring, we randomly sample 40 examples and have the same annotators who originally worked on each example compose a citation-grounded answer from scratch. One of the authors then blindly rates the model- and human-written answers on a five-point Likert scale across answer correctness and completeness. The model-generated answers achieve an average completeness score of 4.80 and an overall quality score of 4.85, closely matching the human-written answers (4.75 for completeness and 4.90 for overall quality). The match indicates that the synthesized references are of sufficiently high quality to serve as reliable targets for LLM-as-Judge scoring.

## E System Prompts Template

Figures 3, 4, 5, and 6 illustrate the prompt templates used in our experiments (agent main prompt, fixed-round response generation, reference-answer generation, and LLM-as-Judge scoring, respectively).

**Deep Research Main Prompt**

Question: {Question}

You are a research agent. Your task is to answer the question by actively using the provided Search Tool. Use the search tool iteratively for many turns. But in each turn, you should only use the search tool once. Refine your queries based on previous results to broaden coverage and fill knowledge gaps. Stop searching only once you have gathered a comprehensive and multi-perspective set of evidence. Your final response must integrate information from different angles, supported by multiple sources. You must base your answer solely on the retrieved evidence documents—do not use any prior or external knowledge.

Your final response should be in the following format:  
 Answer: {{Your final answer. You should cite your evidence documents inline by enclosing their docids in square brackets at the end of sentences. For example, [20].}}  
 Confidence: {{Your confidence score between 0% and 100% for your answer}}

Figure 3: Prompt to run deep research agent.

**Fixed Round Response Generation Prompt**

I will give you a question and a set of evidence documents, which contains helpful information to answer the question.

Question: {Question}

Evidence documents: {EvidenceDocuments}

Your final response must integrate information from different angles, supported by multiple sources. You must base your answer solely on the retrieved evidence documents—do not use any prior or external knowledge.

Your final response should be in the following format:  
 Answer: {{Your final answer. You should cite your evidence documents inline by enclosing their docids in square brackets at the end of sentences. For example, [20].}}  
 Confidence: {{Your confidence score between 0% and 100% for your answer}}

Figure 4: Prompt to generate the final response after a fixed round of retrieval. At each fixed round  $r \in \{1, 2, 3\}$ , {EvidenceDocuments} is the concatenation of all documents retrieved through round  $r$ .

**Reference Answer Generation Prompt**

Your task is to write a single, comprehensive, citation-grounded reference answer to the given query.

**Rules:**

- Use **only** the provided positive passages as evidence.
- Cover **all** listed reasoning aspects; weigh content by the given importance if provided.
- Cite supporting documents inline using square brackets with the provided IDs, e.g., [doc\_12].
- Provide a comprehensive and well-structured final answer with citations.

**[QUERY]**  
 {{query\_text}}

**[REASONING ASPECTS]**  
 {{aspects\_with\_ids\_and\_optional\_weights}}

**[POSITIVE PASSAGES]**  
 (each item includes: id, aspect, and full text)  
 {{positive\_passages\_block}}

**[OUTPUT FORMAT]**  
 Write a single reference answer in paragraphs. Include inline citations such as [doc\_id].

Figure 5: Prompt for reference answer generation, showing input structure and output specification.

### LLM-as-Judge Scoring Prompt (System)

You are an expert evaluator grading a research-assistant's answer.

For each example you receive:

- **QUESTION** — a query from a specialized StackExchange community.
- **REASONING\_ASPECTS** — a list of key sub-questions / premises a correct answer must address. Each aspect has a short id ( $a_1, a_2, \dots$ ). Treat the list as the authoritative rubric.
- **REFERENCE\_ANSWER** — a citation-grounded answer produced from the gold evidence passages. This is the high-quality target.
- **SYSTEM\_ANSWER** — the answer produced by the system under evaluation.

**Step 1 — score EACH aspect on a 3-point coverage scale:**

- 1.0 fully addressed with specific, correctly-supported claims
- 0.5 partially addressed (mentioned but shallow, OR addressed with notable inaccuracies, OR right idea but missing critical detail)
- 0.0 not addressed, off-topic, or factually wrong

You **MUST** grade every aspect id given in REASONING\_ASPECTS — do not skip any.

**Step 2 — assign one holistic overall quality score (1–5 integer)** for SYSTEM\_ANSWER relative to REFERENCE\_ANSWER (correctness, structure, citations, coherence, no hallucinations):

- 5 matches or exceeds REFERENCE\_ANSWER
- 4 slightly worse but still correct and well-structured
- 3 correct but less thorough or less clearly structured
- 2 partially correct with notable issues
- 1 mostly wrong or hallucinated

Return **STRICTLY** a single JSON object, no markdown fences, no prose outside the object:

```
{"aspect_scores": {"a1": <0|0.5|1>, "a2": <0|0.5|1>, ...}, "overall_quality": <int 1-5>, "justification": "<1-2 sentence rationale>"}
```

---

#### User message

QUESTION:

{question}

REASONING\_ASPECTS:

{aspects\_block}

(*bulleted: - a1: <text>, w=<likert>*)

REFERENCE\_ANSWER:

{reference\_answer}

SYSTEM\_ANSWER:

{system\_answer}

Figure 6: Prompt for LLM-as-Judge scoring of system responses. The judge returns one  $\{0, 0.5, 1\}$  coverage score per aspect (`aspect_scores`) plus a 1–5 holistic `overall_quality` score; the 1–5 `reasoning_completeness` reported in the main text is computed post-hoc as  $\text{round}(4 \cdot \bar{w} + 1)$ , where  $\bar{w}$  is the weight-normalized mean of the per-aspect coverage scores using the Likert aspect weights.

## F Agent Decoding and Search Configuration

All agentic experiments share the LLM-side decoding and tool-side search settings listed below across both agent backends, so that performance differences across retrievers are attributable to retrieval quality rather than to generation variability or agent stopping behavior.

**Agent Backends.** We evaluate every retriever with two complementary backends, both queried through an OpenAI Responses-API-compatible interface:

- **GPT-5-mini-08-07**, accessed through the Azure OpenAI Responses API. We set `reasoning_effort=medium` and leave `temperature` and `top_p` at the model’s defaults (not overridden). We allow up to 30,000 output tokens per search/reasoning turn and 10,000 tokens for the final answer.
- **Qwen3.5-122B-A10B-GPTQ-Int4** (Yang et al., 2025a), self-hosted via vLLM v0.19.1 with `-reasoning-parser qwen3` and `-tool-call-parser qwen3_coder`, exposing a stateless `/v1/responses` endpoint. We allow up to 25,600 output tokens per turn and 12,800 tokens for the final answer; `temperature` and `top_p` are left at the served defaults.

Both backends use identical system/user prompts, tool schema, top- $k$  setting, snippet-truncation budget, and round/iteration limits.

**Search Tool.** Each call to the search tool returns the top-5 retrieved documents. The text of each returned document is truncated to 2,048 tokens with the Qwen/Qwen3-0.6B tokenizer before being shown to the agent, so that a single long document cannot dominate the context window. The agent is given a single `search` tool (no separate `get_document` tool); the tool schema and prompt template are shared across backends and across retrievers.

**Interaction Budgets.** Under the *fixed-round* protocol the agent runs the conversation through search rounds  $r \in \{1, 2, 3\}$  and, at each of those rounds, generates a final answer conditioned on all documents retrieved up to round  $r$  (a safety cap of 10 tool-calling iterations is set but is never the binding constraint, since termination is triggered at

round 3). Under the *adaptive-round* protocol the agent decides when to stop on its own, up to a hard limit of 100 rounds.

**Sampling.** For both protocols we evaluate on a fixed sample of 175 queries (25 per task, drawn with a fixed random seed from the 739-query BRIGHT-PRO pool); every retriever is evaluated on the identical sample, so cross-retriever differences are not confounded by question selection.

## G Data Examples

---

**Query:** Did neanderthals need vitamin C from the diet? Modern humans need to get vitamin C from the diet, because we do not make it ourselves. Did neanderthals produce vitamin C? At what point of evolution exactly was this ability lost?

---

**Reasoning Aspect 1 (weight = 0.27)**

Vitamin C biosynthesis in animals requires the enzyme L-gulonolactone oxidase (encoded by the *GULO/GLO* gene), which catalyzes the final step of the pathway.

**Positive Documents**

**Doc 1:** (...abbrev...) Review of vertebrate vitamin C synthesis losses in teleost fishes, anthropoid primates, guinea pigs, and some bats and Passeriformes; in every studied case the inability to synthesize is caused by mutations in the *GLO* gene that codes for the enzyme catalyzing the last step of biosynthesis (the figure of the biosynthesis pathway labels enzyme 6 as L-gulonolactone oxidase, GLO, EC 1.1.3.8). (...abbrev...)

**Doc 2:** (...abbrev...) “Why are mutations limited to *GLO* genes?” — because losing *GLO* only affects vitamin C production, whereas losing genes for other enzymes in the same pathway (e.g. gluconolactonase, EC 3.1.1.17) would also disrupt caprolactam degradation and the pentose phosphate pathway; the *GLO* gene is therefore “predisposed” to being lost since it makes a single compound unnecessary for other pathways. (...abbrev...)

---

**Reasoning Aspect 2 (weight = 0.27)**

Anthropoid primates carry a mutated, nonfunctional *GULOP* pseudogene, preventing endogenous vitamin C synthesis and forcing dietary intake.

**Positive Documents**

**Doc 1:** (...abbrev...) The nonfunctional *GULOP* pseudogene maps to human chromosome 8p21; loss of *GULO* activity occurred independently in some bats, in guinea pigs, and in the haplorrhine suborder of primates (which includes humans). The mutated remnant is still present in guinea pig and human genomes. (...abbrev...)

---

**Reasoning Aspect 3 (weight = 0.27)**

Loss of functional *GULO* in primates occurred about 61–63 million years ago, around the haplorrhine–strepsirrhine split.

**Positive Documents**

**Doc 1:** (...abbrev...) Anthropoid primates have lost 7 of the 12 exons of *GLO*; inactivation dates derived from comparing functional and nonfunctional sequences place the event at ~61 MYA in anthropoid primates and ~14 MYA in guinea pigs, consistent with phylogenetic estimates. (...abbrev...)

---

**Reasoning Aspect 4 (weight = 0.18)**

Neanderthals belong to the haplorrhine hominid lineage and share the degenerated *GULOP* sequence with modern humans, so they too could not synthesize vitamin C.

**Positive Documents**

**Doc 1:** (...abbrev...) Haplorrhini and strepsirrhini primates diverged ~70 MYA; *GULOP* is found in all haplorrhini, while strepsirrhini retain a functional *GULO*. Neanderthal genomic data place them within the haplorrhine hominids alongside *Pongo*, *Gorilla*, *Pan*, and *Homo*. (...abbrev...)

---

Table 6: BRIGHT-PRO Biology Example.

---

**Query:**

I have never understood why earth's inner core is solid. Considering that the inner core is made of an iron–nickel alloy (melting point around 1350 °C to 1600 °C) and the temperature of the inner core is approximately 5430 °C (about the temperature of the surface of the sun). Since Earth's core is nearly 3–4 times the melting point of iron–nickel alloys, how can it possibly be solid?

---

**Reasoning Aspect 1 (weight = 0.30)**

For most materials the melting point rises with pressure: the solid phase is denser than the liquid, so higher temperatures are required to overcome intermolecular forces and melt the material at high pressure.

**Positive Documents**

**Doc 1:** (...abbreviation...) Pressure–temperature phase diagrams show that the melting curve generally has a positive slope for most substances (water being the anomalous exception with a negative slope), so increasing pressure raises the melting temperature. (...abbreviation...)

---

**Reasoning Aspect 2 (weight = 0.30)**

Earth's inner core experiences extreme pressures of ~330–360 GPa, which greatly elevate the melting temperature of iron and iron–nickel alloys above the local geotherm.

**Positive Documents**

**Doc 1:** (...abbreviation...) Wikipedia article on Earth's inner core: pressure ranges from ~330 to 360 GPa (3.3–3.6 million atm); “iron can be solid at such high temperatures only because its melting temperature increases dramatically at pressures of that magnitude (see the Clausius–Clapeyron relation)”; the inner core is solid in accordance with the Simon–Glatzel equation. (...abbreviation...)

**Doc 2:** (...abbreviation...) Description of Earth's outer and inner core: “The only reason why iron and other heavy metals can be solid at such high temperatures is because their melting temperatures dramatically increase at the pressures present there, which ranges from about 330 to 360 gigapascals”; the outer core remains liquid because it “is not under enough pressure to be solid” despite a similar iron–nickel composition. (...abbreviation...)

---

**Reasoning Aspect 3 (weight = 0.20)**

Experimental and theoretical studies estimate iron's melting temperature near inner-core boundary pressures to be roughly 5,500–7,000 K, well above the local temperature.

**Positive Documents**

**Doc 1:** (...abbreviation...) Review article: “Previous studies give a range of iron melting temperatures from 5500 to 7000 K at the ICB pressure” across extrapolations from static laser-heated diamond-anvil-cell measurements (Anzellini et al., 2013; Morard et al., 2018), dynamic shock-wave compressions (Brown & McQueen, 1986; Yoo et al., 1993), and ab initio theoretical calculations (Alfè, 2009; Bouchet et al., 2013). (...abbreviation...)

**Doc 2:** (...abbreviation...) Companion treatment of the iron melting curve at ICB (~330 GPa): same 5500–7000 K range from static, dynamic, and theoretical studies; specific shock-wave determinations include 5600(500) K at ~243 GPa (Brown & McQueen) and 6830(500) K and 7800(500) K at the ICB (Yoo et al.; Bass et al.). (...abbreviation...)

---

**Reasoning Aspect 4 (weight = 0.20)**

The inner core is solid while the outer core is liquid because the pressure-driven melting curve crosses the geotherm at the inner-core boundary.

**Positive Documents**

**Doc 1:** (...abbreviation...) Inner core: ~1220 km radius, iron–nickel alloy, surface temperature ~5430 °C (~ Sun's photosphere); the high-pressure solid phase is inferred from seismic waves and Earth's magnetic field. (...abbreviation...)

**Doc 2:** (...abbreviation...) “The isentropic temperature profile in the liquid iron alloy in the outer core of Earth intersects the iron melting curve at ~330 GPa, forming the outer boundary of the solid inner core” — in other words, the geotherm crosses iron's pressure-elevated melting curve at the ICB, leaving the deeper part solid and the shallower part liquid (where convection drives the geodynamo). (...abbreviation...)

---

Table 7: BRIGHT-PRO Earth Science Example.

---

**Query:**

Is there any example for “pluralistic ignorance” in economics? I read in Bursztyn (2021) that misperceptions about others can be explained by stereotyping, motivated reasoning, and pluralistic ignorance — the situation in which almost all members of a group privately reject group norms, yet believe that virtually all other members accept them. Could you give an example in economics or finance to make this concrete?

---

**Reasoning Aspect 1 (weight = 0.25)**

Pluralistic ignorance arises when individuals privately hold preferences differing from perceived majority norms, leading them to publicly conform despite personal disagreement (“preference falsification”).

**Positive Documents**

**Doc 1:** (...abbreviation...) Following Kuran (1995), “public preference falsification” refers to settings where people publicly express positions opposite to their privately held views; it differs from white lies (it brings disutility) and from self-deception (it is a social phenomenon driven by conformity to perceived majority preferences). (...abbreviation...)

---

**Reasoning Aspect 2 (weight = 0.25)**

In Saudi Arabia, most young married men privately support women working outside the home but underestimate other men’s support, suppressing their wives’ labor-force participation.

**Positive Documents**

**Doc 1:** (...abbreviation...) Bursztyn, González, and Yanagizawa-Drott (2020) document low female labor-force participation in Saudi Arabia (~15%) alongside even lower observable employment of wives (~4% in their sample); using a survey experiment they show that men privately support women working outside the home but systematically underestimate other men’s support, a textbook case of pluralistic ignorance. (...abbreviation...)

---

**Reasoning Aspect 3 (weight = 0.17)**

Tax-compliance research shows many taxpayers personally value honesty yet believe that others tolerate evasion, sustaining a mistaken descriptive norm of noncompliance.

**Positive Documents**

**Doc 1:** (...abbreviation...) Wenzel (2001) and others find that self-reported compliance correlates with beliefs about others’ compliance, and that most people believe one should be honest yet think other taxpayers are not honest — a divergence between injunctive and descriptive norms that fits the pluralistic-ignorance pattern. (...abbreviation...)

---

**Reasoning Aspect 4 (weight = 0.17)**

In studies of bribery, individuals routinely overestimate community approval of corrupt practices even though only small minorities actually endorse them.

**Positive Documents**

**Doc 1:** (...abbreviation...) In a multi-state Nigerian study of pass-mark bribery, respondents in all states except Sokoto systematically overestimated the share of community members who hold positive normative beliefs about bribery: e.g., respondents in Adamawa and Enugu thought ~40% of their community held such beliefs, when in fact only 8.6% and 11% did; the same gap holds for traffic-bribery solicitation (perceived 40–50% vs. actual 8–8.2%). (...abbreviation...)

---

**Reasoning Aspect 5 (weight = 0.17)**

On U.S. climate policy, supporters outnumber opponents two-to-one, yet 80–90% of Americans underestimate national support for major mitigation policies, discouraging advocacy and collective action.

**Positive Documents**

**Doc 1:** (...abbreviation...) Representative U.S. survey ( $N = 6,119$ ) finds a near-universal misperception (“false social reality”): 66–80% of Americans support major climate mitigation policies, but average estimates of public support are only 37–43%; the misperception holds across every state and demographic, illustrating pluralistic ignorance as a barrier to collective action on climate change. (...abbreviation...)

---

Table 8: BRIGHT-PRO Economics Example.

---

**Query:** What is the scientific term for the tendency to see familiar patterns in things that are actually something completely different? One of the most common examples is perceiving human faces in clouds, cars, and all kinds of objects. I'm looking for a low-level misperception that people are usually aware of on a higher cognitive level (i.e., they know they are not actually seeing a face).

---

**Reasoning Aspect 1 (weight = 0.33)**

*Pareidolia* is the perceptual tendency to impose meaningful interpretations on ambiguous (typically visual) stimuli, detecting objects or patterns where none actually exist.

**Positive Documents**

**Doc 1:** (...abbreviation...) Encyclopedic definition: "Pareidolia is the tendency for perception to impose a meaningful interpretation on a nebulous stimulus, usually visual, so that one detects an object, pattern, or meaning where there is none. Pareidolia is a type of apophenia." (...abbreviation...)

**Doc 2:** (...abbreviation...) Same definition expanded with examples and scope: "a specific but common type of apophenia"; common examples include perceived images of animals, faces, or objects in cloud formations, faces in inanimate objects, and lunar pareidolia (Man in the Moon, Moon rabbit); the concept extends to hearing voices or music in random noise (e.g., from air conditioners or fans), and face pareidolia has been demonstrated in rhesus macaques. (...abbreviation...)

---

**Reasoning Aspect 2 (weight = 0.22)**

Pareidolia is a specific subtype of *apophenia* — the broader human tendency to perceive meaningful connections within unrelated data.

**Positive Documents**

**Doc 1:** (...abbreviation...) Apophenia (Klaus Conrad, 1958) is defined as the "unmotivated seeing of connections accompanied by a specific feeling of abnormal meaningfulness"; introduced in the context of early-stage schizophrenia and distinguished from hallucination. (...abbreviation...)

---

**Reasoning Aspect 3 (weight = 0.22)**

Neuroimaging shows face-like (pareidolic) stimuli activate face-processing regions such as the fusiform face area (FFA), similarly to real faces.

**Positive Documents**

**Doc 1:** (...abbreviation...) A 2009 magnetoencephalography study showed that objects perceived as faces evoke an early (~165 ms) FFA activation similar in time and location to that evoked by real faces (~130 ms), supporting face perception of face-like objects as an early process rather than a late cognitive reinterpretation. (...abbreviation...)

**Doc 2:** (...abbreviation...) Kanwisher et al. identified the Fusiform Face Area (FFA) on the lateral mid-fusiform gyrus, with greater fMRI activation to faces than to letter strings or textures; Liu et al. further showed the FFA plays a crucial role in both real-face perception and face-pareidolia processing, with higher FFA activation during face pareidolia than during letter pareidolia, supporting an interaction between bottom-up (occipitotemporal) and top-down (frontal) processing. (...abbreviation...)

---

**Reasoning Aspect 4 (weight = 0.22)**

Brain responses to illusory faces occur rapidly and may initially represent objects as faces before higher-level processing resolves the error.

**Positive Documents**

**Doc 1:** (...abbreviation...) MEG plus fMRI multivariate decoding shows that illusory faces are first represented as more face-like than matched objects in face-selective occipito-temporal cortex, and only ~100 ms later are represented as more object-like — consistent with a rapidly-deployed, broadly-tuned face-detection system whose output is subsequently corrected. (...abbreviation...)

---

Table 9: BRIGHT-PRO Psychology Example.

---

**Query:**

ROS2 — modify Python launchfile without rebuilding package. In ROS1 we could modify XML launchfiles and `roslaunch` them directly without `catkin_make` rebuilding, but in ROS2 it seems if I modify a Python launchfile, I need to rebuild the package; otherwise `ros2 launch <package_name> <launchfile_name>` runs the old launchfile. Do I need to rebuild the package after each modification, or am I missing something?

---

**Reasoning Aspect 1 (weight = 0.30)**

ROS2 launch files are packaged and installed into the package's share directory during build, and `ros2 launch` executes the installed copy — not the source.

**Positive Documents**

**Doc 1:** (... abbreviation ...) ROS2 tutorial “Integrating launch files into ROS2 packages” shows the standard layout (`launch/` directory at package top level) and explains that launch files are installed and consumed from the package share directory after building with `ament_python/ament_cmake`. (... abbreviation ...)

---

**Reasoning Aspect 2 (weight = 0.20)**

By default `colcon` performs out-of-source builds, placing built artifacts in an `install/` directory peer to `src/`, so source edits do not affect what `ros2 launch` runs.

**Positive Documents**

**Doc 1:** (... abbreviation ...) A ROS2 workspace contains `src/`, and `colcon` creates `build/`, `install/`, and `log/` as peers; each package is installed into its own subdirectory of `install/` (no `devel` as in `catkin`). (... abbreviation ...)

---

**Reasoning Aspect 3 (weight = 0.30)**

Building with `colcon build -symlink-install` replaces the install copies with symbolic links pointing back at the source files, so subsequent edits take effect immediately without rebuilding.

**Positive Documents**

**Doc 1:** (... abbreviation ...) `colcon build` verb documentation describing flags such as `-build-base`, `-install-base`, `-symlink-install`, and `-merge-install`, and how install behavior is controlled. (... abbreviation ...)

**Doc 2:** (... abbreviation ...) Community Q&A confirms that with `-symlink-install`, source code changes take effect without recompilation (for Python; not C++). (... abbreviation ...)

---

**Reasoning Aspect 4 (weight = 0.20)**

With `-symlink-install` enabled, modifying a launch file (or any Python file) in `src/` becomes immediately visible to `ros2 launch` without rebuilding the package.

**Positive Documents**

**Doc 1:** (... abbreviation ...) Stack Exchange answer explaining that `-symlink-install` creates symlinks in the `install` directory pointing at the actual source files, so launchfile changes do not require a rebuild. (... abbreviation ...)

---

Table 10: BRIGHT-PRO Robotics Example.

---

**Query:** Lossy compression: `uint16`→`uint8`. I need a lossy compression that maps a `uint16` to a `uint8` so that the resolution loss *increases with the magnitude* of the input. My current attempt is `log2_compress(x) = uint8((log2(x+1)/16)*255)`, which exploits that  $\log_2(1) = 0$  and  $\log_2(65536) = 16$ . This is simple but wasteful on the `uint8` side (e.g., codewords `[1, 14]`, `[16, 24]`, ... are never used). Can someone suggest a method similar to `log2_compress` but that uses (more of) all 256 `uint8` levels?

---

**Reasoning Aspect 1 (weight = 0.38)**

Mapping `uint16` into `uint8` with magnitude-dependent loss corresponds to *nonuniform quantization* (smaller intervals where the signal is dense, larger where it is sparse), not uniform linear quantization.

**Positive Documents**

**Doc 1:** (...abbreviation...) Textbook treatment of nonuniform quantization: by analogy with assigning shorter codewords to high-probability symbols in lossless compression, a quantizer can use smaller intervals in regions of high probability mass and larger intervals where the signal is rare, lowering average distortion. (...abbreviation...)

---

**Reasoning Aspect 2 (weight = 0.38)**

$\mu$ -law companding applies a logarithmic compression that concentrates quantization resolution near small magnitudes while compressing larger values — the standard 8-bit companding scheme used in G.711 PCM telephony.

**Positive Documents**

**Doc 1:** (...abbreviation...) The  $\mu$ -law algorithm (G.711) is a companding algorithm primarily used in 8-bit PCM digital telecommunications; its non-linear quantization effectively increases the dynamic range covered for a given number of bits, at a small cost in peak SNR. (...abbreviation...)

---

**Reasoning Aspect 3 (weight = 0.25)**

The *Lloyd–Max* algorithm designs optimal nonuniform scalar quantizers by iteratively choosing decision thresholds and reconstruction levels to minimize mean-squared quantization error for a given source PDF.

**Positive Documents**

**Doc 1:** (...abbreviation...) Lloyd–Max scalar quantizer: for a signal  $x$  with PDF  $f_X(x)$  and  $M$  representation levels, decision thresholds lie midway between adjacent reconstruction levels, and reconstruction levels lie at the centroid of the PDF between thresholds; the iterative procedure yields the MSE-optimal nonuniform quantizer. (...abbreviation...)

---

Table 11: BRIGHT-PRO Stackoverflow Example.

---

**Query:** Is it okay from a sustainability viewpoint to buy a Christmas tree? I am thinking about buying a Christmas tree but I want to know if that is or isn't good from a sustainability viewpoint.

---

**Reasoning Aspect 1 (weight = 0.33)**

Life-cycle assessment (LCA) comparing natural and artificial Christmas trees generally finds that natural trees have lower annual climate and resource impacts.

**Positive Documents**

**Doc 1:** (...abbreviation...) ISO 14040/14044 LCA study (Montreal market) compares a plantation natural tree harvested 150 km south of Montreal with a 6-year-reuse artificial tree manufactured in China and shipped via Vancouver; on an annual basis the artificial tree has  $\sim 3\times$  more impacts on climate change and resource depletion than the natural tree ( $\sim 8.1$  vs.  $\sim 3.1$  kg CO<sub>2</sub>/yr); natural tree wins on climate change and resources but loses on ecosystem quality (mainly land occupation). (...abbreviation...)

---

**Reasoning Aspect 2 (weight = 0.22)**

Artificial trees concentrate environmental impacts in manufacturing and long-distance transport, becoming climate-favorable only if reused for many years.

**Positive Documents**

**Doc 1:** (...abbreviation...) Carbon Trust guidance: the real-vs-artificial question has no one-size-fits-all answer; "an artificial tree, used over multiple years (7–20 years, depending on weight and materials), is better for reducing emissions than buying a new commercially grown tree every year," though plastic/metal components may end up as waste. (...abbreviation...)

---

**Reasoning Aspect 3 (weight = 0.22)**

Sustainability of real Christmas trees depends on responsible forestry, with third-party certification systems verifying that supplying forests are managed to sustainability standards.

**Positive Documents**

**Doc 1:** (...abbreviation...) The American Tree Farm System (ATFS), the United States' oldest family-forest certification program, applies eight sustainability standards verified by independent third-party audits; certified family forests are managed for water, wildlife, wood, and recreation. (...abbreviation...)

**Doc 2:** (...abbreviation...) Forest certification is defined as the certification of land-management practices to a standard of sustainability, attested by an independent third party. (...abbreviation...)

---

**Reasoning Aspect 4 (weight = 0.22)**

Real Christmas trees are farm-grown over several years, can provide wildlife habitat while growing, and can be recycled into mulch or compost at end of life.

**Positive Documents**

**Doc 1:** (...abbreviation...) Christmas tree plantations function like managed woodland meadows, providing habitat for songbirds, ground-dwelling birds, mammals, and pollinators via diverse ground covers between trees. (...abbreviation...)

**Doc 2:** (...abbreviation...) The U.S. produces Christmas trees on farms across many states (Oregon and North Carolina lead with  $\sim 4.7$ M and  $\sim 4$ M trees/year); typical Christmas trees are conifers with 600+ species available, grown over multiple years before harvest. (...abbreviation...)

**Doc 3:** (...abbreviation...) Municipal recycling programs accept natural Christmas trees for conversion into mulch and compost (decorations and flocking removed), saving landfill space; flocked trees cannot be recycled. (...abbreviation...)

---

Table 12: BRIGHT-PRO Sustainable Living Example.

## H Experiment Results

Model	BRIGHT Overall	Biology	Earth Science	Economics	Psychology	Robotics	Stack Overflow	Sustainable Living	Overall
BGE-Reasoner-8B	33.8	<u>91.0</u>	<b>91.4</b>	<u>82.7</u>	<b>84.0</b>	<b>82.0</b>	<b>86.0</b>	<b>83.9</b>	<b>85.9</b>
DIVER-4B-1020	30.6	<b>92.1</b>	87.6	<b>83.2</b>	<u>82.2</u>	<u>75.1</u>	<u>83.8</u>	<u>77.7</u>	<u>83.1</u>
DIVER-4B	28.9	87.9	89.6	74.6	80.9	69.6	79.3	74.5	79.5
RTriever-4B (ours)	27.7	84.4	83.5	75.9	71.5	64.6	76.8	75.3	76.0
INF-Retriever-Pro (7B)	26.3	83.5	80.6	74.8	74.0	65.1	70.4	71.9	74.3
Qwen3-8B	23.7	70.9	77.5	64.8	66.0	61.2	72.1	59.3	67.4
INSTRUCTOR-XL (1.5B)	18.9	65.9	74.8	61.3	66.0	62.1	65.7	63.8	65.7
GTE-7B	22.5	79.2	78.7	61.0	60.9	51.7	59.9	56.9	64.0
OpenAI-Embed-3L	17.9	72.8	75.1	63.5	65.8	50.3	60.3	59.6	63.9
GritLM (7B)	21.0	69.3	74.3	59.9	57.6	56.8	58.0	57.4	61.9
ReasonIR-8B	24.4	63.6	69.5	54.9	54.3	56.4	68.7	53.4	60.1
BM25	14.5	58.3	67.6	57.5	44.6	53.7	60.6	59.6	57.4
EmbeddingGemma-300M	18.9	68.6	76.5	49.5	51.2	45.9	53.4	54.7	57.1

Table 13: A-Recall@25 across domains ( $\times 100$ ). BRIGHT Overall reports the corresponding NDCG@10 on the original BRIGHT dataset.

Model	BRIGHT Overall	Biology	Earth Science	Economics	Psychology	Robotics	Stack Overflow	Sustainable Living	Overall
BGE-Reasoner-8B	33.8	<u>80.3</u>	<b>84.0</b>	<u>71.0</u>	<b>74.4</b>	<b>71.8</b>	<b>82.1</b>	<b>68.9</b>	<b>76.1</b>
DIVER-4B-1020	30.6	<b>81.4</b>	79.0	<b>71.3</b>	<u>71.9</u>	<u>62.1</u>	<u>79.5</u>	<u>62.7</u>	<u>72.5</u>
DIVER-4B	28.9	77.0	<u>79.7</u>	60.5	71.0	57.3	<u>74.1</u>	57.3	68.1
RTriever-4B (ours)	27.7	69.6	71.1	60.5	60.7	53.5	71.5	57.1	63.4
INF-Retriever-Pro (7B)	26.3	70.0	69.2	59.3	64.9	53.5	65.5	53.9	62.3
Qwen3-8B	23.7	57.2	64.3	50.0	56.9	50.4	66.3	42.5	55.4
INSTRUCTOR-XL (1.5B)	18.9	52.5	62.8	49.7	54.1	49.4	59.6	45.6	53.4
GTE-7B	22.5	66.1	68.1	45.3	49.1	40.5	53.8	40.4	51.9
OpenAI-Embed-3L	17.9	56.9	60.1	48.8	54.7	38.4	54.9	42.9	51.0
GritLM (7B)	21.0	55.2	62.3	45.2	48.2	45.6	52.2	39.9	49.8
ReasonIR-8B	24.4	47.3	56.4	41.2	44.0	45.9	62.7	38.0	47.9
BM25	14.5	45.1	53.6	44.9	35.9	43.8	53.6	43.4	45.8
EmbeddingGemma-300M	18.9	54.2	63.3	37.7	41.5	35.7	46.0	39.6	45.4

Table 14: Recall@25 across domains ( $\times 100$ ). BRIGHT Overall reports the corresponding NDCG@10 on the original BRIGHT dataset.

Model	BRIGHT Overall	Biology	Earth Science	Economics	Psychology	Robotics	Stack Overflow	Sustainable Living	Overall
BGE-Reasoner-8B	33.8	<b>73.3</b>	<b>75.5</b>	<b>64.1</b>	<b>64.7</b>	<b>63.0</b>	<b>68.8</b>	<b>61.7</b>	<b>67.3</b>
DIVER-4B-1020	30.6	<u>72.4</u>	<u>72.9</u>	<u>60.4</u>	59.0	<u>54.0</u>	<u>64.4</u>	<u>56.9</u>	<u>62.9</u>
DIVER-4B	28.9	67.8	71.5	51.4	<u>59.6</u>	50.9	60.8	50.8	59.0
RTriever-4B (ours)	27.7	61.7	63.4	51.3	48.6	46.7	53.4	50.6	53.7
INF-Retriever-Pro (7B)	26.3	60.8	62.8	48.3	51.3	44.0	50.3	47.3	52.1
Qwen3-8B	23.7	51.2	57.5	42.8	45.5	45.6	52.5	38.2	47.6
INSTRUCTOR-XL (1.5B)	18.9	43.7	53.6	42.6	43.9	42.4	47.6	38.6	44.6
OpenAI-Embed-3L	17.9	51.7	52.6	42.0	43.2	35.4	44.1	37.9	43.8
GTE-7B	22.5	58.7	61.4	35.9	36.2	39.2	39.7	34.2	43.6
GritLM (7B)	21.0	48.7	56.7	37.5	36.3	41.3	38.4	35.8	42.1
ReasonIR-8B	24.4	40.1	50.1	32.6	33.7	38.6	45.7	32.2	39.0
BM25	14.5	40.2	47.2	38.1	29.6	37.6	39.6	37.9	38.6
EmbeddingGemma-300M	18.9	47.3	55.1	30.5	31.7	32.7	34.8	32.8	37.8

Table 15: NDCG@25 across domains ( $\times 100$ ). BRIGHT Overall reports the corresponding NDCG@10 on the original BRIGHT dataset.

Model	Round 1			Round 2			Round 3		
	$\alpha$ -nDCG (0-100)	Compl. (1-5)	Overall (1-5)	$\alpha$ -nDCG (0-100)	Compl. (1-5)	Overall (1-5)	$\alpha$ -nDCG (0-100)	Compl. (1-5)	Overall (1-5)
BGE-Reasoner-8B	<b>56.36</b>	<b>3.72</b>	<b>3.49</b>	<b>59.50</b>	<b>3.99</b>	<b>3.77</b>	<b>61.85</b>	<b>4.08</b>	<b>3.89</b>
GTE-7B	41.86	3.36	3.17	46.69	3.76	3.61	<u>49.78</u>	3.89	<u>3.72</u>
RTriever-4B (ours)	<u>43.92</u>	<u>3.45</u>	<u>3.26</u>	<u>46.94</u>	<u>3.81</u>	<u>3.63</u>	49.65	3.93	3.71
DIVER-4B	40.02	3.33	3.19	44.60	3.78	3.56	47.93	<u>3.94</u>	<u>3.73</u>
Qwen3-8B	38.74	3.23	3.06	43.28	3.69	3.45	46.46	3.86	3.64
DIVER-4B-1020	36.11	3.17	2.98	40.52	3.65	3.43	43.55	3.88	3.66
GritLM (7B)	34.40	3.12	3.01	38.23	3.55	3.41	41.89	3.85	3.58
BM25	32.98	3.08	2.95	38.51	3.67	3.48	41.31	3.87	3.66
INSTRUCTOR-XL (1.5B)	29.60	3.10	2.88	35.86	3.53	3.33	38.65	3.74	3.52
ReasonIR-8B	30.98	3.12	2.98	35.57	3.47	3.29	37.97	3.65	3.43

Table 16: Performance of retrievers under BRIGHT-PRO fixed-round agentic evaluation setting with the Qwen3.5-122B-A10B agent (counterpart of Table 3, which uses GPT-5-mini). Because each round returns top-5 passages, we report  $\alpha$ -nDCG at the cumulative cut-off:  $k=5$  after round 1,  $k=10$  after round 2, and  $k=15$  after round 3. **Compl.**: reasoning completeness of the final response; **Overall**: overall quality of the final response. Rows are ordered by round-3  $\alpha$ -nDCG@15 (descending).

Case Study (Early-Round Efficiency): Why is the Antarctic ice sheet only a few kilometres thick?

Aspect coverage by retrieval (3 rounds total)

Aspect (weight)	Gold retrieved
<b>Outward gravitational flow</b> (3): snowfall accumulates in the interior but ice continually flows outward toward the coast.	2 / 3
<b>Self-limiting basal stress</b> (2): increasing thickness raises basal stress and accelerates deformation, so ice cannot grow indefinitely.	2 / 2
<b>Mass-balance closure</b> (3): interior accumulation is offset by ablation, basal melting under shelves, calving, and peripheral discharge.	4 / 5
<b>Slow-flow domes</b> (2): near central domes and divides ice velocities are very small, allowing ancient ice and subglacial features to persist.	3 / 3

Per-round search queries issued by the agent

- **R1.** “Dome Fuji 25 mm/yr accumulation . . . oldest ice ice flow basal melting ice divides compression densification”
- **R2.** “Why Antarctic ice thickness limited 3000 m equilibrium between accumulation and ice flow basal melt geothermal heat. . .”
- **R3.** “Nature 1.5 million year ice core Antarctica oldest ice Vostok Dome Fuji”

**Final retrieval:** 11 / 13 gold; **judge:** weighted aspect coverage = 1.00, overall quality = 5.

RTriever Response (excerpt)

*“Antarctic snow does not simply pile up indefinitely in one place; most of the mass that falls in the interior is continuously redistributed and removed by ice flow, basal melting/freezing and surface ablation, and the ice sheet finds a dynamic equilibrium thickness (kilometres, not hundreds of kilometres) set by those processes and by ice rheology. . . Ice sheets are not static piles: they flow outward from high domes toward the coast; ice at the summit is fed by snowfall but the ice mass is transported laterally by deformation and sliding so net thickness is controlled by the balance between input (snowfall) and output (flow, basal melt, calving, sublimation). . . Thicker ice and increased surface slope raise driving stress, which increases deformation and flow; glaciers and ice sheets therefore thicken until flow (and other losses) balances accumulation, producing steady-state thicknesses of a few kilometres rather than hundreds of kilometres.”*

**What the evidence shows.** The first two rounds retrieve ten of the eleven gold documents, covering all four reasoning aspects, and round 3 only adds a peripheral hit on slow-flow domes; the agent then terminates. Because high-quality evidence is concentrated in the first two queries, the LLM never has to speculate about why ice does not pile up indefinitely. It composes a mass-balance answer that cites the gold passages on outward flow, basal stress, ablation, and dome dynamics directly. This is the canonical pattern RTriever shows on broadly-aspected scientific queries: rapid early concentration of gold evidence, mass-balance reasoning paid out aspect-by-aspect, and a tight three-round termination.

Figure 7: RTriever on “Why is the Antarctic ice sheet only a few kilometres thick?” (EARTH\_SCIENCE qid=48). Three search rounds suffice to cover all four reasoning aspects.

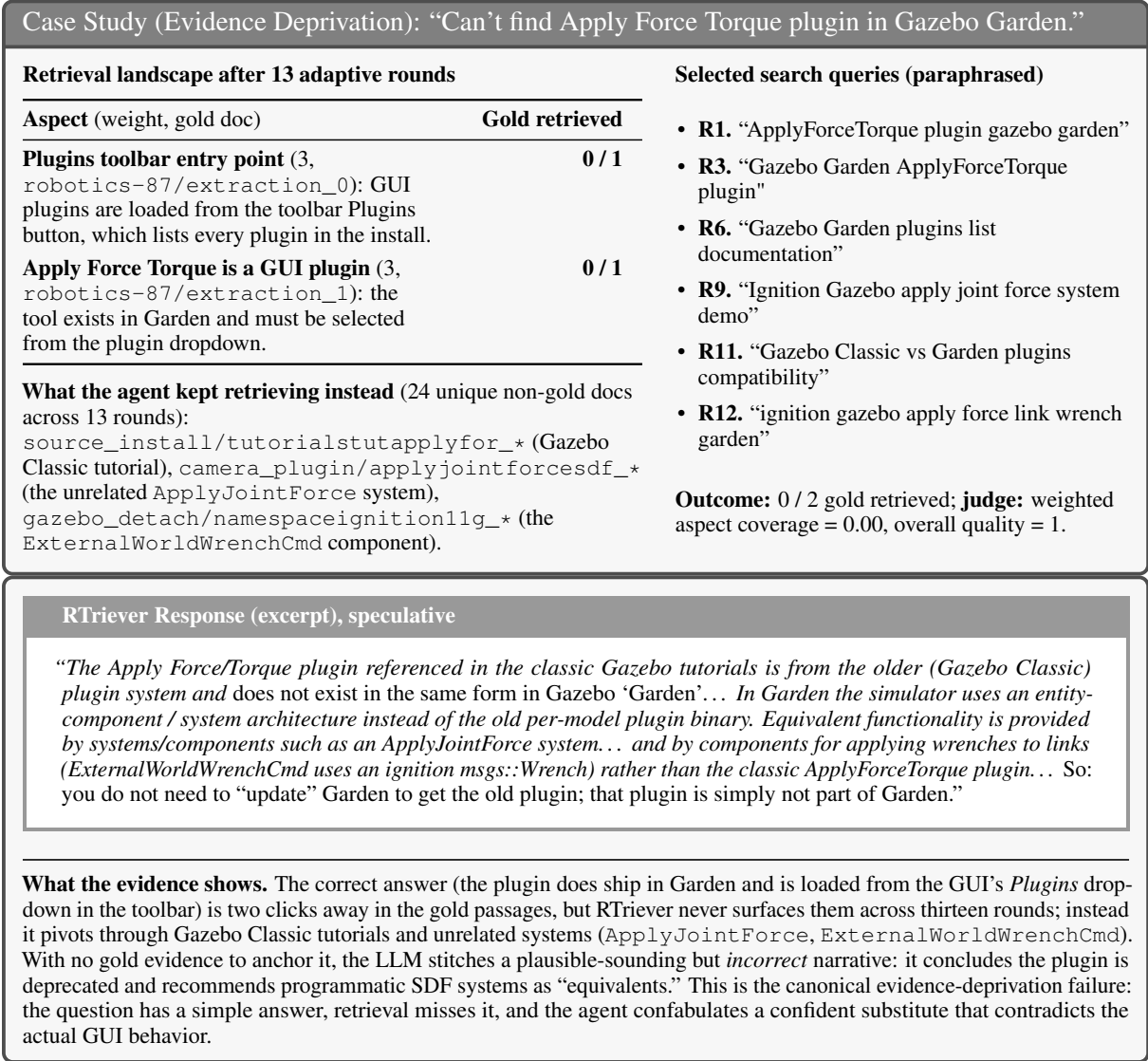


Figure 8: RTriever on “Can’t find Apply Force Torque plugin in Gazebo Garden” (ROBOTICS qid=87). Thirteen rounds, zero gold retrieved; the model speculates that the plugin no longer exists.

Case Study (Repetition Bias): Has interfamilial hybridization ever succeeded?

Documents repeatedly resurfaced across 12 rounds

Document ID	Type	Times
Primate_3_14 (closest non-primate relatives)	non-gold / off-topic	7
Human_81_1 (human monogamy)	non-gold / off-topic	5
Human_14_483 (human monogamy)	non-gold / off-topic	3
Human_14_127 (human monogamy)	non-gold / off-topic	3
biology-64d3516332ed	non-gold / off-topic	3
Parthenogenesis_4_5	non-gold / tangential	3
Parthenogenesis_4_2	non-gold / tangential	3
<b>Subtotal (non-gold repeats)</b>		<b>27</b>
biology-86/extraction_4	gold (1 of 7)	4
Reproductive_isolation_2	gold (1 of 7)	3
<b>Subtotal (gold repeats)</b>		<b>7</b>

What the agent never reaches

- Reproductive\_isolation\_0/\_1 (pre-zygotic barriers; Aspect 1, weight 3)
- biology-86/extraction\_0 (post-zygotic isolation; Aspect 2, weight 3)
- biology-86/extraction\_3 (interfamilial fish hybrids; Aspect 3, weight 3)

The four *Ursus* hybrid documents and the daylily-Lycoris embryo-rescue passage are reached, but only after rounds 3 and 6; the rest of the run circles back to the same primate / human-monogamy / parthenogenesis cluster.

**Aggregate:** 60 retrieval slots → 28 unique documents → 32 cross-round duplicates (≈53% of the slot budget); **gold:** 3 / 7 retrieved (extraction\_4, extraction\_5, Reproductive\_isolation\_2); **judge:** weighted aspect coverage = 0.59, overall quality = 3.

RTriever Response (excerpt), partial coverage

*“In animals: basically no credible examples . . . of true extra-familial sexual hybrids that produced viable, reproducing offspring; what is documented are interspecific and intergeneric hybrids within the same family or closely related genera. . . In plants: extra-familial hybrids are rare but have been produced under artificial (laboratory/ breeding) conditions using assisted methods such as embryo rescue; at least one documented case produced a plantlet identified as an interfamily hybrid. . . A reported study crossed daylily (Hemerocallis) with Lycoris and, after many aborted ovaries and using assisted techniques, recovered one embryo that developed into a plantlet identified as an interfamily hybrid (L. aurea × H. ‘My Complementary’). . . The literature emphasizes that hybrid formation yielding fertile, stable offspring is uncommon in animals because post-zygotic barriers (chromosomal mismatches, gene incompatibilities) prevent viable, fertile hybrids as genetic distance increases.”*

**What the evidence shows.** The first round latches onto Primate\_3\_14 and the parthenogenesis cluster; over the next eleven rounds the agent rephrases its query seven different ways (“intergeneric hybridization,” “hybrids between different families,” “somatic hybridization protoplast fusion,” “gelada hamadryas baboon hybrid,” . . .) but the retriever resurfaces the same anchors plus an unrelated human-monogamy cluster, producing 32 cross-round duplicate retrievals on a 60-slot budget. The agent does eventually find three of the seven gold documents, but the two highest-weighted aspects (*pre-zygotic barriers* and *post-zygotic isolation*, both weight 3) are never reached. The final response correctly describes the daylily-Lycoris example but treats reproductive-isolation mechanisms only at a generic level, yielding the partial answer that the aspect-aware judge scores at  $wac = 0.59$ . The failure mode is search-dynamics: feedback from the early rounds narrows rather than expands the candidate set, even when the agent is actively trying to broaden the query.

Figure 9: RTriever on “Is sexual reproduction outside the same biological family possible?” (BIOLOGY qid=86). Repeated off-target retrievals dominate twelve adaptive rounds.

Case Study (Aspect Tunnel Vision): GHCN climate data; what is EMXT 317, and how do I get the bulk data?

Per-aspect retrieval coverage (7 adaptive rounds)

Aspect (weight, gold doc)	Retr.
a1. EMXT denotes extreme maximum temperature, derived from daily TMAX ( <code>ncdc-meta</code> ).	✓
a2. Climate-division datasets provide bulk monthly temperature/precipitation by division/county, often in CSV ( <code>earth-sci-98/ext_0,_1</code> ).	✗
a3. The NCEI Access Data Service API enables HTTP queries returning CSV ( <code>earth-sci-98/ext_2</code> ).	✗

The user explicitly asked two things: (i) what does the value 317 mean, and (ii) is there a better way to get U.S. monthly climate data from 1992–2012. The agent saturates on (i) and never queries any concept relevant to (ii).

All seven search queries (paraphrased)

- **R1.** GHCN monthly EMXT meaning, code, units, “317”
- **R2.** GHCN Monthly CSV units tenths degrees EMXT
- **R3.** GHCN-Daily TMAX units tenths format
- **R4.** GHCN-Monthly EMXT 317 interpretation, scale
- **R5.** GHCN-Daily “tenths of degrees,” TMAX
- **R6.** EMXT CSV sample, Petersburg ND
- **R7.** ghcnd file format TMAX tenths

**Judge:** aspect a1 score = 1.00; a2 = a3 = 0; weighted aspect coverage = 0.50, overall quality = 3.

RTriever Response (excerpt), detailed but lop-sided

*“EMXT is the ‘extreme maximum temperature’ for that station-month, i.e., the single highest daily maximum temperature within the month. . . The CSV/text export is the route to get machine-readable, multi-station data. . . How you can get the data you need (state or county, 1992–2012) in CSV form: (1) Use the GHCN Monthly Summaries Custom ASCII Spreadsheet (CSV) export. . . (2) The Monthly Summaries are derived from the GHCN-Daily data; if you need different aggregation rules you can download GHCN-Daily and compute monthly/areal aggregates yourself. . . ”*

**What the evidence shows.** Every one of the seven queries elaborates the same sub-question (“what does EMXT mean / what units”). The user’s other half (which dataset to use for bulk 1992–2012 county-level data) is never queried, so the climate-division dataset and the NCEI Access Data Service API are entirely absent from the retrieval set. The final answer is technically correct on the EMXT semantics and even produces a partial workflow recommendation, but it improvises that workflow on top of the GHCN-Monthly Summaries product (the only one the agent has retrieved evidence for) rather than directing the user to the dataset family designed for their use case. Aspect-aware evaluation surfaces the failure:  $a_2$  and  $a_3$  receive zero coverage, and weighted aspect coverage drops to 0.50 despite a fluent, well-cited response. The pattern of fluent answer combined with unbalanced retrieval is distinct from repetition bias because the search queries themselves *are* novel each round; the lock-in is semantic, not document-level.

Figure 10: RTriever on a two-part GHCN climate-data question (EARTH\_SCIENCE qid=98). All seven search rounds drill into one aspect; the user’s data-source question is never queried.

Case Study (Hypothesis Hopping): Is there a name for “perceived repeated interruption”?

**Round-by-round behaviour**

Round	Search hypothesis	New gold (and aspect)
R1	“perceived repeated interruption / single stall”	ext_1 (a1 frequency illusion)
R2	“Baader–Meinhof phenomenon, confirmation bias”	ext_4 (a4), ext_6 (a5)
R3	spotlight effect, heightened self-awareness	none (non-gold)
R4	illusion of control, “being followed”	none (non-gold)
R5	Jungian synchronicity, meaningful coincidences	none (re-hits ext_6)
R6	clustering illusion, patterns in random events	ext_5 (a4)

**Outcome statistics**

- Rounds: 6 (vs. task median 5)
  - Gold retrieved: 4 / 8 (ext\_1, 4, 5, 6)
  - First gold round: 1; last *new* gold round: 6
  - Aspect scores:  $a_1 = a_2 = a_3 = a_4 = 1, a_5 = 0.5$
  - Weighted aspect coverage = 0.95, overall quality = 4
- The retrieved set ends up dominated by 5 *spotlight\_effect* passages and 5 *illusion\_of\_control* passages, neither tied to a gold-supported aspect.

After round 2 the agent already has a directly-named answer (the *frequency illusion / Baader–Meinhof phenomenon*), plus supporting evidence on availability and patternicity. Rounds 3–6 enumerate alternative concept labels rather than terminating.

**RTriever Response (excerpt), correct but bloated**

*“There isn’t a single agreed-upon name for the exact feeling you describe, but a number of well-documented cognitive biases and concepts cover it; most directly the ‘frequency illusion’ (also called the Baader–Meinhof phenomenon) and the general tendency to perceive patterns in randomness often labeled apophenia or patternicity. . . Other closely related bias names. . . illusory correlation and the clustering illusion. . . Two additional, related concepts people often invoke: the spotlight effect (feeling that others are paying more attention to you than they actually are) can increase self-consciousness about interruptions but is conceptually different. . . Jungian ‘synchronicity’ is a non-empirical label. . . The illusion of control can also play a role. . .”*

**What the evidence shows.** Unlike the early-round case where convergence is rewarded, here *the agent already has the answer* after round 2 but cannot stop. Each subsequent round tests a fresh hypothesis (*spotlight effect, illusion of control, synchronicity, clustering illusion*) before terminating. The final answer remains correct (judge oq = 4, weighted aspect coverage = 0.95), but the retrieval set is diluted with ten passages on the spotlight effect and illusion of control that play no causal role in the answer, and the response itself reads as a survey rather than a direct identification. This pattern is the converse of evidence deprivation: retrieval succeeds early, but lacking a confidence signal, the agent keeps spending rounds. From a deployment standpoint the cost is real, since each extra round adds an LLM call and a retrieval pass, yet it is invisible to final-round-only metrics.

Figure 11: RTriever on “Is there a name for ‘perceived repeated interruption’?” (PSYCHOLOGY qid=98). Gold answer is retrieved by round 2; rounds 3–6 enumerate alternative concept names.