

# Rethinking Evaluation of Reasoning-Intensive Retrieval: Toward Real-World Deep Research Workflows

Anonymous ACL submission

## Abstract

Reasoning-intensive retrieval aims to surface evidence that maximizes downstream reasoning utility rather than only topical similarity. This capability is increasingly vital for agentic retriever-in-the-loop systems such as Deep-Research. However, existing evaluations, exemplified by BRIGHT, provide narrow gold sets and evaluate retrievers in isolation, which obscures their value inside realistic agent workflows. We introduce **BRIGHT-PRO**, an evaluation framework that aligns reasoning-intensive retrieval with deep-research workflows. It extends each query’s positive documents through expert annotation and organizes passages into reasoning aspects, which represent complementary perspectives and subproblems. This structure enables fine-grained analysis of aspect coverage. Beyond static evaluation, we further assess retrievers in realistic deep-research contexts, measuring their practical utility when serving as core components within agentic workflows. Using **BRIGHT-PRO**, we evaluate classical lexical, general-purpose, and reasoning-intensive retrievers, providing actionable insights for developing the next generation of reasoning-intensive retrieval systems.

## 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., 2025c; Weller et al., 2024; Lee et al., 2024). However, as user queries grow more complex, retrieval systems must move beyond surface-level relevance, giving rise to reasoning-intensive retrieval, a new frontier in IR research (SU et al., 2025; Shao et al., 2025; Xiao et al., 2024). Unlike conventional retrieval

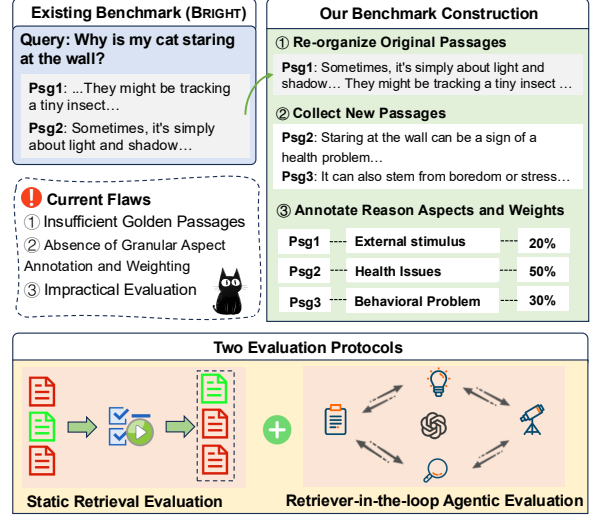


Figure 1: An overview of the **BRIGHT-PRO** benchmark and two evaluation protocols.

that prioritizes surface-level relevance, reasoning-intensive retrieval focuses on finding the most useful pieces of evidence for downstream reasoning. This paradigm shift is crucial for enabling AI systems to handle research-style or analytical queries.

On the other hand, *Deep-Research* represents an emerging agentic paradigm that extends beyond static retrieval (Chen et al., 2025; Du et al., 2025; Yang et al., 2025b; Wu et al., 2025). It enables an LLM-based agent to iteratively plan, search, read, evaluate, and synthesize information, mimicking how human experts gather and analyze information during in-depth investigations (OpenAI, 2025; Perplexity, 2025; Tongyi, 2025). Within this process, the retriever plays a pivotal role. A retriever capable of handling reasoning-intensive queries can (1) construct a balanced, multi-faceted evidence portfolio to support comprehensive reasoning, and (2) guide the agent toward faster convergence, thereby reducing computational overhead.

Despite the growing interest in reasoning-intensive retrieval, current evaluation methods

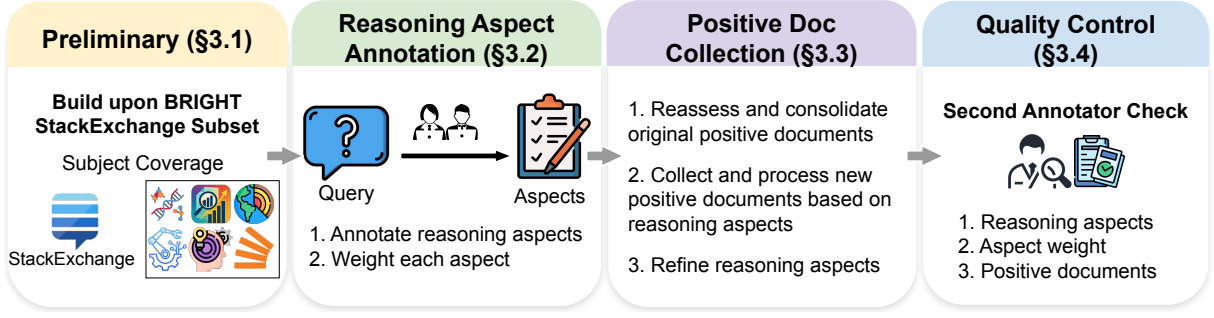


Figure 2: An overview of the **BRIGHT-PRO** benchmark construction pipeline.

remain limited. The only 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. This narrow coverage constrains the benchmark’s ability to assess how effectively a retriever can assemble a broad, multi-faceted evidence set to support complex and comprehensive reasoning. Moreover, the benchmark evaluates retrievers in isolation rather than within the dynamic loop of a deep-research system, thereby failing to capture their true utility in realistic LLM-based agentic workflows.

To bridge this gap, we propose a new evaluation framework, **BRIGHT-PRO**<sup>1</sup>, that extends reasoning-intensive retrieval toward deep-research alignment. Building on BRIGHT, we engage expert annotators to expand each query’s gold passage set and to group these passages into *reasoning aspects* that reflect distinct perspectives or subproblems within the same query, as illustrated in Figure 1. This structure enables fine-grained analysis of whether a retriever achieves comprehensive coverage across multiple reasoning dimensions. Beyond static evaluation, we further assess retrievers in a deep-research context by integrating them into an LLM-based agentic workflow that iteratively plans, retrieves, and synthesizes information. This setup allows us to measure retrievers on system-level outcomes such as reasoning completeness, iteration efficiency, and overall response quality.

We conduct a comprehensive evaluation of retriever systems on **BRIGHT-PRO**, covering classical lexical, general-purpose, and reasoning-intensive retrievers. Across different subsets and evaluation settings, DIVER (Long et al., 2025) consistently delivers the highest performance, while Qwen-family embedders (Li et al., 2023a) form

a strong second tier. Although BM25 performs poorly in the static retrieval setting, it becomes competitive under the agentic evaluation protocol, where the agent’s targeted follow-up queries reduce vocabulary mismatch and allow simple term matching to surface high-utility evidence, demonstrating that classical lexical signals remain valuable when retrieval is embedded in iterative reasoning. Case studies show that early access to gold evidence shortens interaction traces and improves reasoning completeness; in contrast, weaker retrievers invite speculative reasoning when evidence is missing and exhibit repetition bias that recycles low-utility passages rather than discovering complementary aspects. Overall, our results reveal that evaluating retrievers within realistic, iterative research workflows is essential for understanding their true reasoning utility and guiding the development of next-generation reasoning-intensive retrievers.

## 2 BRIGHT-PRO Benchmark

This section introduces the construction of the proposed **BRIGHT-PRO** benchmark, which builds upon BRIGHT (SU et al., 2025) to enable a more comprehensive and realistic evaluation of reasoning-intensive retrieval. Figure 2 outlines the benchmark construction process.

### 2.1 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

<sup>1</sup>Our code and data are available at <https://anonymous.4open.science/r/Bright-Pro>, and will be released publicly upon publication.

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.

## 2.2 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 are required to think beyond the community answers and perform independent analysis to surface the full set of premises needed for a correct and defensible solution. Each aspect is 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 B). We then normalize the scores to weights  $w_a \in [0, 1]$  with  $\sum_a w_a = 1$ .

## 2.3 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<sup>2</sup>. 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

<sup>2</sup>In addition to irrelevant positive passages, we also identified several other problematic cases, which are categorized and discussed in Appendix A. These examples were revised during the data annotation process.

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.

## 2.4 Data Quality Control

To ensure high data quality, each example in **BRIGHT-PRO** is independently re-examined by a second annotator from the same field. Specifically, 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 represent 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. To evaluate the reliability of aspect weighting, we randomly sample 50 queries and ask independent reviewers to rescore the annotated reasoning aspects. The resulting weighted Cohen’s  $\kappa$  of 0.742 suggests that the importance ratings are stable across annotators. After verifying the reasoning aspects, the reviewers are required to 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.

Table 1 presents the key statistics of the **BRIGHT-PRO** benchmark. Examples for each subset can be found in Appendix D.

## 3 BRIGHT-PRO Evaluation Protocol

We evaluate the reasoning-intensive retrieval capabilities of retrievers along two complementary

Subset	Total Number				Avg. Length			Example
	Q	$\mathcal{D}$	$\mathcal{D}^+$	$\mathcal{A}$	Q	$\mathcal{D}$	$\mathcal{A}$	
Biology	103	59,683	9.7	3.8	115.2	130.9	19.6	Tab. 6
Earth Science	116	123,732	9.0	3.7	109.5	154.5	21.0	Tab. 7
Economics	103	52,411	10.0	3.6	181.5	164.7	18.7	Tab. 8
Psychology	101	54,913	9.5	3.4	149.6	159.3	18.4	Tab. 9
Robotics	101	63,999	7.2	3.4	818.9	155.8	21.6	Tab. 10
Stack Overflow	117	109,260	5.4	3.1	478.3	713.8	20.1	Tab. 11
Sustain. Living	108	63,410	12.4	3.6	148.5	148.9	18.5	Tab. 12

Table 1: Data statistics of **BRIGHT-PRO**. For each subset, we show the number of queries ( $Q$ ) and documents ( $\mathcal{D}$ ), the average number of positive documents ( $\mathcal{D}^+$ ) and aspects ( $\mathcal{A}$ ) per example, and the average tokens of queries, documents, and aspects (measured by the GPT-2 tokenizer).

dimensions: a *static* setting, which isolates retrieval quality using aspect-annotated gold passage, and an *agentic* setting, which assesses the retriever’s utility when integrated into a deep-research workflow.

### 3.1 Static Retrieval Evaluation

We first evaluate retrievers in a traditional IR setting, where the system receives a query and returns a single ranked list of passages. As discussed in Section 2.2, 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  $C_j(k) = \sum_{r=1}^k \text{rel}_{rj}$  be the count of retrieved gold items for aspect  $j$  within the top  $k$ . We define binary relevance  $\text{rel}_r = \mathbb{1}\{d_r \in \mathcal{G}\}$ , then the aspect-specific relevance is

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

**$\alpha$ -nDCG@k.** We apply  $\alpha$ -nDCG@ $k$  (Clarke et al., 2008) as the primary metric for evaluation. This metric explicitly models aspect-level diversity, ensuring that retrievers are rewarded for covering multiple reasoning aspects rather than repeatedly retrieving redundant passages about the same subtopic. Specifically, the gain at rank  $r$  is

$$G(r) = \sum_{j=1}^m w_j \text{rel}_{rj} (1 - \alpha)^{C_j(r-1)} \quad (2)$$

where  $\alpha \in [0, 1]$  represents the penalization factor that controls the degree of novelty. In practice,



we set  $\alpha = 0.5$ . We accumulate with logarithmic discount to obtain

$$\text{DCG}_{\alpha}@k = \sum_{r=1}^k \frac{G(r)}{\log_2(r+1)} \quad (3)$$

For normalization, we compute  $\text{IDCG}_{\alpha}@k$  on the gold pool using the same gain definition and a greedy maximization over  $k$  positions. The final metric is defined as

$$\alpha\text{-nDCG}@k = \frac{\text{DCG}_{\alpha}@k}{\text{IDCG}_{\alpha}@k} \quad (4)$$

**Weighted Aspect Recall.** To directly capture coverage of reasoning aspects, we also report a weighted aspect recall. The metric credits an aspect once it is covered at least once:

$$\text{A-Recall}@k = \sum_{j=1}^m w_j \cdot \mathbb{1}\{C_j(k) \geq 1\}, \quad (5)$$

We also use  $\text{nDCG}@k$  and  $\text{Recall}@k$  metrics as complementary metrics.

### 3.2 Experiment Setup for Agentic Evaluation

To evaluate the practical utility of retrievers within a deep-research workflow, we design an agentic evaluation protocol. This setup keeps the LLM-based agentic pipeline fixed while varying only the retrievers across experiments. The current subsection details the experimental setup, and the next subsection describes the evaluation settings.

**Implementation of Agentic Retriever-in-the-Loop System.** To perform LLM-based agentic search, we equip the LLM, *i.e.*, GPT-5-mini-08-07<sup>3</sup>, with a search tool that enables it to query the document corpus and retrieve relevant materials across multiple rounds of interaction. For each retriever, we construct and index a corresponding document corpus. At the initial turn, we provide the query and explicitly prompt the LLM to engage in multi-round information seeking using the available tools. During each iteration, the agent is instructed to issue a new search query, inspect the retrieved evidence, and determine whether additional retrieval is necessary before synthesizing a final response. The model is explicitly instructed to

<sup>3</sup>In our preliminary experiment, we find that frontier open-source models such as Qwen3-Next-80B-A3B-Instruct (Yang et al., 2025a) generally fail to follow tool-use instructions reliably in multi-turn tool-usage settings.

terminate the search once it judges that the accumulated evidence is sufficient to answer the question with high confidence. To ensure transparency and traceability, the agent is required to include explicit citations to the supporting documents used in its final answer. The complete prompting setup is provided in Appendix C.

### LLM-as-Judge for System Response Evaluation.

Following prior work on evaluating DeepResearch agents (Chen et al., 2025), we adopt the LLM-as-Judge framework (Li et al., 2024) with GPT-5 as evaluator for system response evaluation, along two dimensions: (1) *reasoning completeness*, which measures the extent to which the response fully captures the reasoning aspects and evidence needed to answer the question, and (2) *overall quality*, which measures the overall quality of the answer.

To construct reference answers for each query, we use GPT-5 configured with a high reasoning effort setting. Specifically, the model receives human-annotated reasoning aspects along with the full content of the positive passages linked to each aspect and is prompted to generate a comprehensive, citation-grounded answer for each question. To evaluate the quality of the generated answers, we randomly sample 20 examples. For each sampled example, the same annotators who previously worked on that example are asked to compose a citation-grounded answer. One of the authors then blindly rates both the model- and human-generated answers on a five-point Likert scale, assessing 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-annotated answers (4.75 for completeness and 4.90 for overall quality). These results demonstrate that the generated answers are of sufficiently high quality to serve as reliable reference answers. Given the query, system-generated answer, and reference answer, the LLM-as-Judge system evaluates each response and assigns a Likert score from 1 to 5 for both reasoning completeness and overall quality.

### 3.3 Agentic Evaluation Settings

We next discuss the two settings (*i.e.*, fixed-round and adaptive-round) for agentic evaluation.

**Setting 1: Fixed-Round Evaluation.** In this setup, the Deep-Research agent performs a pre-defined number of search rounds (1, 2, and 3). In each round, the system returns the top-5 retrieved

Model	BRIGHT Overall	Biology	Earth Science	Economics	Psychology	Robotics	Stack Overflow	Sustainable Living	Overall
DIVER-4B	<b>28.9</b>	<b>66.1</b>	<b>71.3</b>	<b>54.6</b>	<b>61.6</b>	<b>56.7</b>	<b>58.7</b>	<b>57.2</b>	<b>60.9</b>
INSTRUCTOR-XL	18.9	44.3	55.1	<u>44.9</u>	<u>47.4</u>	46.1	47.4	<u>44.1</u>	<u>47.1</u>
GTE-7B	22.5	<u>58.9</u>	<u>61.6</u>	39.4	40.5	43.8	40.1	38.9	46.2
Qwen3-8B	23.7	<u>38.9</u>	<u>55.6</u>	41.8	45.1	<u>50.1</u>	<u>48.5</u>	39.8	45.7
OpenAI-Embed-3L	17.9	46.2	52.0	40.2	42.9	41.4	44.3	43.7	44.4
GritLM	21.0	50.0	56.7	40.5	39.1	45.7	38.2	40.5	44.4
ReasonIR-8B	<u>24.4</u>	42.4	51.6	35.9	36.9	43.7	44.9	38.0	41.9
BM25	14.5	42.3	48.7	40.3	31.6	42.0	40.9	41.7	41.1

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.

passages. After the first, second, and third rounds, the agent generates a response conditioned on all documents retrieved up to that point. This setup enables a fine-grained, round-by-round analysis, allowing controlled comparison across retrievers under identical interaction budgets. By fixing the number of search rounds, we ensure fair evaluation, attributing performance differences solely to retriever quality rather than to agent stopping behavior or varying search depth. We evaluate the (1)  $\alpha$ -nDCG@k of the retrieved passages and (2) the **reasoning completeness** and **overall quality** of the final responses by LLM-as-Judge system.

**Setting 2: Adaptive-Round Evaluation.** In this setup, the LLM agent decides when to terminate search based on internal confidence and the perceived sufficiency of retrieved evidence. Unlike the fixed-round setting, the adaptive protocol reveals how well a retriever tracks the agent’s evolving information intent. When the retriever fails to surface what the LLM is seeking, the agent issues additional queries and the interaction trace lengthens. Conversely, an effective retriever quickly provides high-leverage evidence that satisfies the agent’s implicit goals, which enables early termination. The number of retrieval rounds therefore serves as an emergent signal of intent alignment and evidence sufficiency. Adaptive-round evaluation assesses both final answer quality and interaction efficiency, and highlights whether a retriever can anticipate rather than only react to the agent’s needs. We report the **iteration count** and the **reasoning completeness** and **overall quality** of the model’s reasoning. Moreover, to quantify the trade-off between overall quality and efficiency, we define the **Efficiency-Quality Reward (AER)**, which balances response quality with the number of retrieval

iterations:

$$\text{AER} = OQ \times e^{-\gamma(R-1)} \quad (6)$$

where  $OQ$  is the overall quality of the response,  $R$  denotes the total number of search rounds executed by the agent, and  $\gamma$  is a tunable decay parameter controlling how strongly later rounds are penalized. We set  $\gamma$  to be 0.05.

## 4 Experiment

This section introduces our experimental results and analysis on the **BRIGHT-PRO** benchmark.

### 4.1 Evaluated Models

We conduct a comprehensive evaluation of various retrieval models, encompassing three major categories: (1) *Classical Lexical Baselines*: **BM25** (Robertson et al., 2009), the canonical sparse lexical retriever that matches queries to documents based on term frequency and inverse document frequency statistics. (2) *General-Purpose Retrievers*: **GritLM** (Muennighoff et al., 2024), a generative-representational instruction-tuned model that unifies text generation and embeddings in a single LLM. **INSTRUCTOR** (Su et al., 2022), an instruction-finetuned embedder that concatenates task instructions with text so a single model can tailor embeddings across many tasks and domains without further finetuning. **GTE** (Li et al., 2023b), a general-purpose embedding Qwen family trained with multi-stage contrastive learning over large, diverse corpora. **Qwen3-Embedding** (Zhang et al., 2025b), a dense embedding retriever built on the Qwen3 foundation model family. **OpenAI Text Embedding Large** (Neelakantan et al., 2022), OpenAI’s most capable general-purpose embedding model, designed for high-accuracy semantic search and

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)
DIVER-4B	<b>50.5</b>	<b>3.62</b>	<b>4.06</b>	<b>53.1</b>	<b>3.93</b>	<b>4.18</b>	<b>55.9</b>	<b>4.09</b>	<b>4.23</b>
GTE-7B	49.9	3.56	4.03	53.0	3.91	4.15	55.6	4.06	4.21
GritLM	43.7	3.32	3.91	47.0	3.44	4.05	49.9	3.61	4.08
INSTRUCTOR-XL	43.0	3.19	3.77	46.6	3.47	3.98	49.8	3.63	4.04
BM25	40.7	3.15	3.75	43.6	3.28	3.93	46.2	3.47	4.00
ReasonIR-8B	39.3	3.02	3.64	43.2	3.17	3.80	45.9	3.44	3.94

Table 3: Performance of retrievers under **BRIGHT-PRO** fixed-round agentic evaluation setting. **Compl**: reasoning completeness of the final response; **Overall**: overall quality of the final response.

clustering. (3) *Reasoning-Intensive Retrievers*: **ReasonIR** (Shao et al., 2025), a retriever explicitly trained for reasoning tasks using a synthetic pipeline that pairs each document with a challenging query and a hard negative. **DIVER** (Long et al., 2025), a dense bi-encoder retriever fine-tuned on synthetic, reasoning-heavy data across domains with carefully constructed hard negatives.

For each model, we employ its native tokenizer, embedding interface, and pooling strategy without any fine-tuning. Input sequences are truncated to each model’s validated maximum context length. For those 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 experiments are conducted on a server equipped with four NVIDIA A100-80GB GPUs.

## 4.2 Main Findings

**Comparing retrievers under static retrieval setting.** Table 2 reports the  $\alpha$ -nDCG@ $k$  performance of the retrievers on **BRIGHT-PRO**, while detailed A-Recall@ $k$ , nDCG@ $k$ , and Recall@ $k$  scores are provided in Appendix E. Across all subsets, DIVER consistently achieves the highest scores, highlighting the effectiveness of its domain-specific synthetic data generation strategy. In contrast, the performance of other retrievers varies across subsets. For example, INSTRUCTOR-XL ranks second on the economics, psychology, and sustainable living subsets but performs relatively poorly on the biology subset. Notably, although ReasonIR achieves competitive results on **BRIGHT**, it performs considerably worse on **BRIGHT-PRO**, underscoring the importance of generalization.

**Comparing retrievers under fixed-round setting.** Table 3 shows that DIVER-4B leads from the first interaction and maintains its advantage across rounds on all three axes, with small but con-

Model	# Round	Compl.	Overall	AER
DIVER-4B	<b>4.31</b>	<b>4.23</b>	<b>4.47</b>	<b>3.81</b>
GTE-7B	5.72	4.12	4.26	3.39
INSTRUCTOR-XL	6.04	3.83	4.06	3.16
BM25	7.37	3.92	4.19	3.02
GritLM	7.13	3.95	4.12	3.01
ReasonIR-8B	7.15	3.89	4.03	2.98

Table 4: Performance of retrievers under **BRIGHT-PRO** adaptive-round agentic evaluation setting.

sistent gaps over GTE-7B in  $\alpha$ -nDCG, reasoning completeness, and overall quality. All systems improve monotonically from Round 1 to Round 3, which indicates that the agent’s follow-up queries partially compensate for early misses. Notably, although weaker under a static retrieval setting, the lexical BM25 baseline proves simple yet effective in the agentic setup: it closely tracks the neural GritLM by Round 3 in  $\alpha$ -nDCG (49.8 vs. 49.9) and even surpasses ReasonIR-8B and INSTRUCTOR-XL across all rounds. This result suggests that straightforward term-matching can still surface high-utility evidence when the agent’s queries are specific and vocabulary-aligned, highlighting that traditional retrieval signals remain valuable under LLM-based agentic control.

**Comparing retrievers under adaptive-round setting.** Table 4 reveals differences that fixed-round views would blur. DIVER-4B attains the highest overall quality while requiring the fewest rounds on average, yielding the best AER. GTE-7B is a close second in overall quality but needs more iterations, which lowers its reward. BM25 again demonstrates its simplicity and effectiveness: despite a longer interaction trace, it achieves higher final overall quality than GritLM (4.19 vs. 4.12). However, its lower AER reflects reduced efficiency due to additional retrieval rounds. These results indicate that while BM25 remains competitive in end quality, frontier reasoning-intensive retrievers such as DIVER-4B

excel in aligning with the agent’s evolving information intent, delivering high-leverage evidence earlier and enabling more efficient reasoning.

### 4.3 Qualitative Analysis

To better assess the impact of retrievers in agentic workflows, we conduct an in-depth qualitative analysis under the adaptive-round setting, using examples from DIVER and ReasonIR. Our analysis reveals the following key findings.

**Early-Round Retrieval Efficiency.** Retrievers that successfully retrieved gold documents in early rounds typically require fewer iterations to achieve comprehensive reasoning coverage. As shown in Figure 7, early access to high-quality evidence enables the LLM to converge more rapidly toward accurate and complete explanations, minimizing redundant or exploratory queries in later rounds.

**Evidence Deprivation and Speculative Reasoning.** Models with weaker retrieval performance often produce responses grounded in assumptions or incomplete reasoning rather than verifiable evidence, as shown in Figure 6. When gold documents are absent, the LLM tend to overgeneralize or infer missing links from loosely related passages, leading to speculative and incorrect explanations.

**Repetition Bias and Self-Reinforcing Retrieval Errors.** Weak retrievers exhibit a persistent *repetition bias*: once early candidates are surfaced, the system keeps reselecting the same few passages—whether off-topic, low-utility, or only partially relevant—instead of discovering new evidence. As shown in Figure 8, the model repeatedly retrieves non-gold or tangential documents across rounds, exhausting the same limited evidence pool. This fixation amplifies early mistakes: initial misfires monopolize later retrievals, starving the search of fresh signals and freezing the evidence frontier.

## 5 Related Work

**Reasoning-Intensive Information Retrieval.** Beyond traditional keyword- and semantic-based IR, BRIGHT (SU et al., 2025) is the first—and so far the only—benchmark that explicitly targets queries requiring multi-step reasoning to identify genuinely useful evidence rather than merely superficial relevance. 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., 2025a; Long et al., 2025; Shao et al., 2025). However, current evaluations typically score a retriever against a small pool of gold passages tied to one or two sources, which limits assessment of coverage across complementary reasoning aspects.

**Deep Research Agent and Evaluation.** In parallel, Deep-Research agents combine LLM planning with iterative search, reading, and synthesis to tackle complex queries (Chen et al., 2025; Du et al., 2025; Yang et al., 2025b; Wu et al., 2025). To enable controlled comparisons of agent components, BROWSECOMP-PLUS (Chen et al., 2025) 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 provides limited visibility into how a retriever shapes an agent’s evidence portfolio, iteration budget, and final reasoning quality. We extend reasoning-intensive retrieval toward deep-research alignment in two ways. First, we expand each BRIGHT query with a larger set of human-curated gold passages and organize them into *reasoning aspects*, which enables fine-grained analysis of aspect coverage and balance. Second, we embed retrievers inside an LLM-driven research loop that plans, retrieves, and synthesizes, allowing system-level evaluation on reasoning completeness, iteration efficiency, and final answer quality.

## 6 Conclusion

We introduce the BRIGHT-PRO benchmark to advance reasoning-intensive retrieval toward closer alignment with deep research workflows. BRIGHT-PRO expands each query’s set of positive documents through expert curation, ensuring broader and more complete coverage of all reasoning-relevant evidence. In addition, it annotates these documents with reasoning aspects and proportional importance weights to enable fine-grained evaluation of multi-aspect reasoning coverage. Beyond static retrieval evaluation, it further assesses models in an agentic retriever-in-the-loop setting, measuring their real-world effectiveness as core components within deep research workflows. We envision BRIGHT-PRO as a step toward more holistic evaluation of retrieval models, fostering future IR research on dynamic, reasoning-aligned retrieval and its integration with deep-research workflows.



## Limitations

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 to enhance representativeness. Additionally, the number of examples in BRIGHT-PRO remains limited due to the high cost of human expert annotation. To address this, future work could explore semi-automated or hybrid annotation pipelines to scale up data creation while maintaining quality and reliability.

## References

- Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. [Lit-search: A retrieval benchmark for scientific literature search](#).
- Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, et al. 2025. 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](#).
- 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*.
- Ehsan Kamalloo, Nandan Thakur, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, and Jimmy Lin. 2024. [Resources for brewing beir: 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](#).
- 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](#).
- 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](#).
- OpenAI. 2025. [Introducing deep research](#).
- Perplexity. 2025. [Introducing perplexity deep research](#).
- Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- 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. 2025. [ReasonIR: Training retrievers for reasoning tasks](#). In *Second Conference on Language Modeling*.
- 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](#).
- 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*.

750	Hongjin SU, Howard Yen, Mengzhou Xia, Weijia Shi,	embedding and reranking through foundation models.	807
751	Niklas Muennighoff, Han yu Wang, Liu Haisu, Quan	<i>arXiv preprint arXiv:2506.05176</i> .	808
752	Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jin-		
753	sung Yoon, Serkan O Arik, Danqi Chen, and Tao Yu.	Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang,	809
754	2025. <a href="#">BRIGHT: A realistic and challenging bench-</a>	Huan Lin, Baosong Yang, Pengjun Xie, An Yang,	810
755	<a href="#">mark for reasoning-intensive retrieval</a> . In <i>The Thir-</i>	Dayiheng Liu, Junyang Lin, et al. 2025c. Qwen3	811
756	<i>teenth International Conference on Learning Repre-</i>	embedding: Advancing text embedding and rerank-	812
757	<i>sentations</i> .	ing through foundation models. <i>arXiv preprint</i>	813
		<i>arXiv:2506.05176</i> .	814
758	Tongyi. 2025. Tongyi deepresearch: A new era of open-		
759	source ai researchers. <a href="https://github.com/Alibaba-NLP/DeepResearch">https://github.com/</a>		
760	<a href="https://github.com/Alibaba-NLP/DeepResearch">Alibaba-NLP/DeepResearch</a> .		
761	Orion Weller, Kathryn Ricci, Eugene Yang, Andrew		
762	Yates, Dawn Lawrie, and Benjamin Van Durme. 2025.		
763	<a href="#">Rank1: Test-time compute for reranking in informa-</a>		
764	<a href="#">tion retrieval</a> .		
765	Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ash-		
766	win Paranjape, Yuhao Zhang, and Jack Hessel. 2024.		
767	Promptriever: Instruction-trained retrievers can be		
768	prompted like language models. <i>arXiv preprint</i>		
769	<i>arXiv:2409.11136</i> .		
770	Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang,		
771	Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He,		
772	Deyu Zhou, Pengjun Xie, et al. 2025. Webwalker:		
773	Benchmarking llms in web traversal. <i>arXiv preprint</i>		
774	<i>arXiv:2501.07572</i> .		
775	Chenghao Xiao, G Thomas Hudson, and Noura Al		
776	Moubayed. 2024. Rar-b: Reasoning as retrieval		
777	benchmark. <i>arXiv preprint arXiv:2404.06347</i> .		
778	An Yang, Anfeng Li, Baosong Yang, Beichen Zhang,		
779	Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,		
780	Chengen Huang, Chenxu Lv, Chujie Zheng, Dayi-		
781	heng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge,		
782	Haoran Wei, Huan Lin, Jialong Tang, Jian Yang,		
783	Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi		
784	Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai		
785	Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao		
786	Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang,		
787	Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan		
788	Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao		
789	Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xu-		
790	ancheng Ren, Yang Fan, Yang Su, Yichang Zhang,		
791	Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang,		
792	Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan		
793	Qiu. 2025a. <a href="#">Qwen3 technical report</a> .		
794	John Yang, Kilian Lieret, Carlos E. Jimenez, Alexan-		
795	der Wettig, Kabir Khandpur, Yanzhe Zhang, Binyuan		
796	Hui, Ofir Press, Ludwig Schmidt, and Diyi Yang.		
797	2025b. <a href="#">Swe-smith: Scaling data for software engi-</a>		
798	<a href="#">neering agents</a> .		
799	Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan,		
800	Anh Tuan Luu, and Chen Zhao. 2025a. Diffusion vs.		
801	autoregressive language models: A text embedding		
802	perspective. <i>arXiv preprint arXiv:2505.15045</i> .		
803	Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang,		
804	Huan Lin, Baosong Yang, Pengjun Xie, An Yang,		
805	Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren		
806	Zhou. 2025b. Qwen3 embedding: Advancing text		

## A Types of Problematic Cases Identified in BRIGHT Benchmark

**Irrelevant Gold Documents.** A substantial portion of BRIGHT’s positive documents are only loosely related to their corresponding queries or even completely irrelevant. Instead of providing direct or reasoning-supportive evidence, many contain tangential discussions or unrelated contextual text.

**Incomprehensive Gold Document.** The dataset lacks comprehensive positive gold documents for each query and thus fail to capture the full reasoning context needed to answer the query. Consequently, BRIGHT cannot reliably support retrieval evaluations for deep research since the available gold passages provide only partial reasoning support rather than a comprehensive answer basis.

**Severe Structural Noise.** BRIGHT inherits typical web-crawled artifacts such as HTML tags, navigation bars, and boilerplate segments. These distort the text structure and make the dataset unsuitable for evaluating fine-grained retrieval accuracy.

**Semantic Discontinuity.** Document segmentation in BRIGHT often breaks semantically coherent sentences or paragraphs, resulting in fragmented units that disrupt the logical flow and compromise retrieval evaluation consistency.

## B 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.

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.



## C System Prompts Template

850

Figure 3, 4 illustrates the prompt templates used in our experiments.

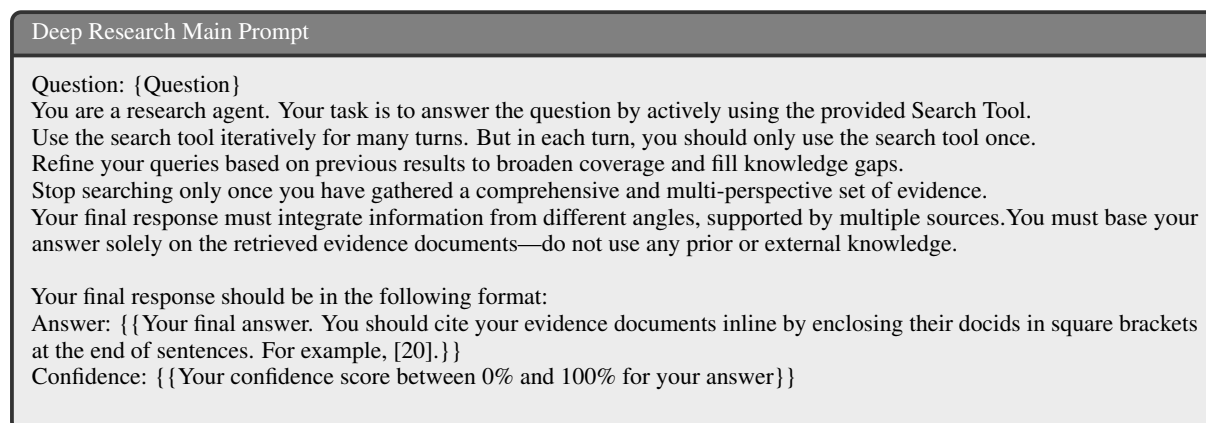


Figure 3: Prompt to run deep research agent

851

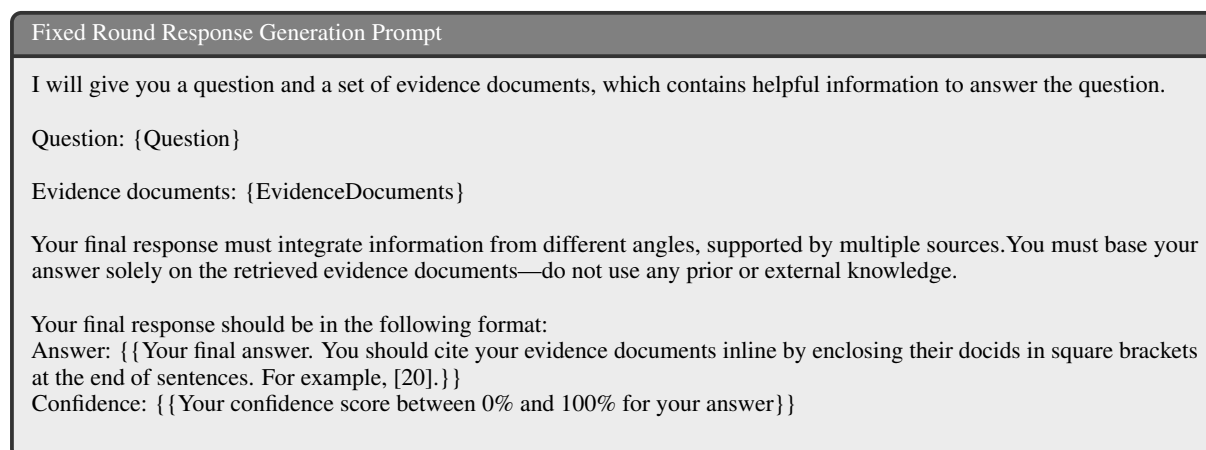


Figure 4: Prompt to generate final response after fixed round of retrieval

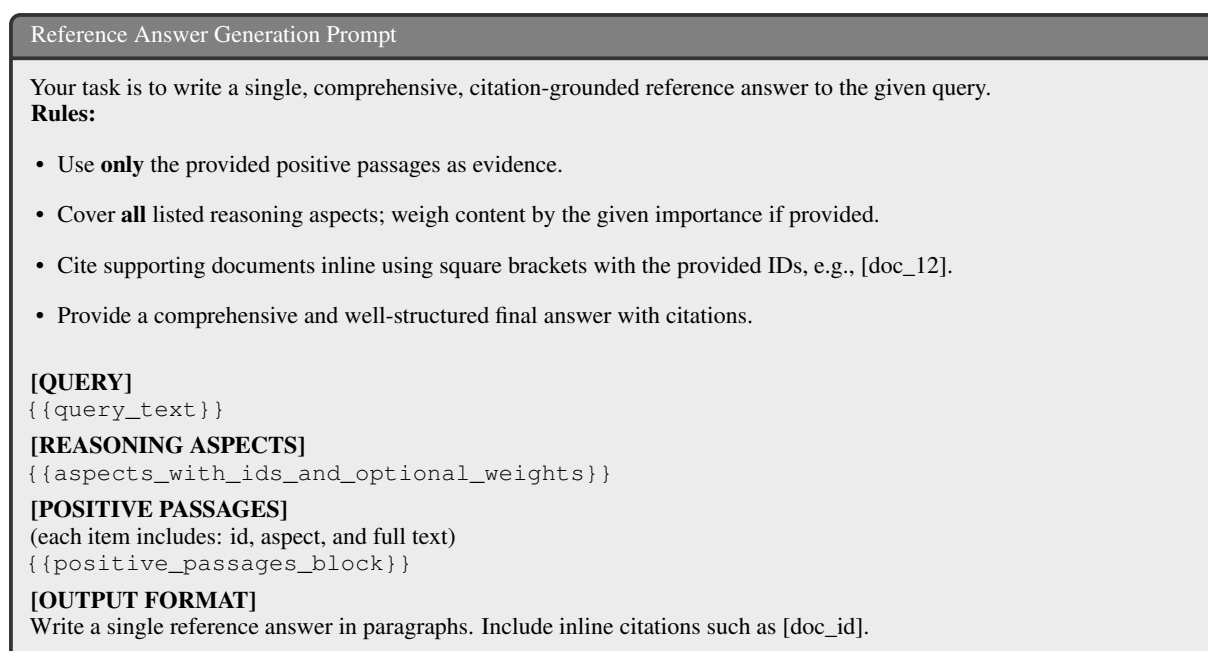


Figure 5: Prompt for reference answer generation, showing input structure and output specification.

---

**Query:** Claim in article about why insects are attracted to light. The article argues: “Heat radiation as an attractive component is refuted by the effect of LED lighting, which supplies negligible infrared radiation yet still entraps vast numbers of insects.” Why would attraction to LEDs imply they are *not* seeking heat? Could insects be evolutionarily programmed to associate light with heat—so they still “expect” heat near LEDs even if it’s not there?

---

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

Artificial lights disrupt insects’ innate visual orientation system, causing them to become trapped as they try to keep their backs aligned with the light source.

**Positive Documents**

**Doc 1:** (...abbrev...) Insects orient their dorsum toward the brightest hemisphere (dorsal-light-response), yielding orthogonal flight, orbiting, stalls, and trapping near point sources; model and 3D kinematics support DLR as the most plausible mechanism. (...abbrev...)

---

**Reasoning Aspect 2 (weight = 0.40)**

Field evidence shows attraction is better explained by spectral composition (especially UV/blue) and brightness than by heat output.

**Positive Documents**

**Doc 1:** (...abbrev...) LEDs caught fewer insects than CFL and filament; UV content strongly predicts attraction. Some thermal effects may occur in specific contexts, but spectral differences dominate overall patterns. (...abbrev...)

**Doc 2:** (...abbrev...) Across street-light types, shorter wavelengths and intensity increase attraction; LEDs emit little short-wavelength light and often attract fewer insects; warmer (3000–4000K) LEDs can further reduce attraction. (...abbrev...)

---

**Reasoning Aspect 3 (weight = 0.10)**

Heat is a strong cue for specific insects (e.g., mosquitoes) during host-seeking, which is a different behavior from the general flight-to-light phenomenon.

**Positive Documents**

**Doc 1:** (...abbrev...) Female *Aedes* mosquitoes, activated by CO<sub>2</sub>, are attracted to heat and humidity; thermal cues guide short-range host finding and interact with visual contrast—distinct from generic light-attraction dynamics. (...abbrev...)

---

**Reasoning Aspect 4 (weight = 0.25)**

Many insects are particularly sensitive to short-wavelength light (UV/blue), which enhances attraction to certain artificial light sources.

**Positive Documents**

**Doc 1:** (...abbrev...) Ancestral UV–blue–green opsins are widespread; positive phototaxis to short wavelengths is common; light-trap studies show disproportionate attraction to short wavelengths. (...abbrev...)

**Doc 2:** (...abbrev...) Most insects peak in UV sensitivity (350–400 nm); navigation via celestial light and polarization cues; dorsal-tilting near artificial lights observed in high-speed recordings. (...abbrev...)

**Doc 3:** (...abbrev...) In mosquitoes, shorter bands (<500 nm) trigger positive phototaxis; long bands (>500 nm) inhibit; narrower-band blue/green LEDs can out-attract brighter white mixes, highlighting wavelength-specific responses. (...abbrev...)

---

Table 6: BRIGHT-PRO Biology Example.

---

**Query:**

We know (think?) that Earth has three Hadley cells per hemisphere, but from observing gas giants such as Jupiter, we see that they have many more cells. According to a link from a comment in this question, Titan might have one cell going from north to south poles. What factors affect the number of cells a planet has, and how? Or to put it another way, given a hypothetical planet with atmosphere, what facts would you need to know in order to guess at how many cells it has?

---

**Reasoning Aspect 1 (weight = 0.5)**

A planet's rotation rate and size are fundamental factors: faster rotation (and/or larger radius) confines Hadley cells toward the equator and encourages multiple midlatitude eddy-driven jets, yielding more, narrower cells; slower rotation permits broad, near-global Hadley cells and fewer meridional cells.

**Positive Documents**

**Doc 1:** (...abbreviation...) For Earth/Mars-like planets, the transition between tropical ( $R_o \gtrsim 1$ ) and extratropical ( $R_o \ll 1$ ) regimes occurs near  $20^\circ$ – $30^\circ$ ; with longer rotation periods, planets become “all tropics” (near-global Hadley) when the period  $\gtrsim 10$  Earth days. (...abbreviation...)

**Doc 2:** (...abbreviation...) In an angular-momentum-conserving upper branch,  $u = \Omega a \frac{\sin^2 \phi}{\cos \phi}$  grows rapidly with latitude; realistic GCMs show that increasing rotation rate confines the Hadley cell and strengthens subtropical jets, while slower rotation yields cells extending to high latitudes. (...abbreviation...)

**Doc 3:** (...abbreviation...) GCM suites varying rotation from  $1/16$  to  $4 \times$  Earth show: slow rotation  $\Rightarrow$  nearly global Hadley cells (few cells); fast rotation  $\Rightarrow$  confined Hadley cells plus extratropical baroclinic zones and multiple jets (more cells). (...abbreviation...)

---

**Reasoning Aspect 2 (weight = 0.2)**

On rapidly rotating planets (e.g., gas giants), the number and spacing of east–west jets relate to the Rhines scale, which depends on characteristic wind speed, the  $\beta$ -effect (rotation/planet size), and thus sets how many jets (and associated cells) can fit meridionally.

**Positive Documents**

**Doc 1:** (...abbreviation...) The  $\beta$  term anisotropizes turbulence into zonal jets; the Rhines scale  $L_R$  characterizes the largest energy-containing structures and relates to jet spacing across regimes. (...abbreviation...)

**Doc 2:** (...abbreviation...) In idealized GCMs, the separation scale between jets is similar to the Rossby radius; larger planets or faster rotation (smaller  $L_R$ ) lead to more jets per hemisphere. Jovian jet spacings match this scaling. (...abbreviation...)

**Doc 3:** (...abbreviation...) Shallow-water and QG studies with small-scale forcing on rapidly rotating spheres produce multiple zonal jets whose meridional width scales close to  $(U/\beta)^{1/2}$  (Rhines). Forcing/damping choices modulate equatorial jet direction but multiple jets persist. (...abbreviation...)

---

**Reasoning Aspect 3 (weight = 0.3)**

Heating pattern, seasonality (obliquity), atmospheric mass/stratification, and radiative timescales shape cell width/strength and can merge/split cells (e.g., near-global single cells on slowly rotating or tidally locked worlds versus two- or three-cell patterns on Earth-like cases).

**Positive Documents**

**Doc 1:** (...abbreviation...) Hadley cells are thermally direct responses to equator–pole insolation gradients; their width scales with parameters like  $\Delta\theta$ , tropopause height  $H_t$ , rotation rate  $\Omega$ , and radius  $a$  (e.g., Held–Hou scaling). (...abbreviation...)

**Doc 2:** (...abbreviation...) Seasonality shifts the ascending branch with the subsolar latitude: near equinox two similar cells; near solstice a stronger cross-equatorial cell forms—configuration depends on atmospheric heat capacity and radiative damping timescales. (...abbreviation...)

**Doc 3:** (...abbreviation...) Atmospheric structure (static stability, deformation radius) and moist processes affect eddy scales and jet spacing; smaller Rossby radii or stronger baroclinicity yield more jets, while large radiative timescales favor broad overturning. (...abbreviation...)

**Doc 4:** (...abbreviation...) Detailed Hadley-cell mechanics and diagnostics: definition via zonally averaged streamfunction; seasonal transition between two symmetric cells (equinox) and a dominant cross-equatorial cell (solstice); Held–Hou width scaling  $\phi \propto \sqrt{\frac{g \Delta\theta H_t}{\Omega^2 a^2 \theta_0}}$  highlights roles of insolation contrast, tropopause height, rotation rate, and radius in setting cell extent. (...abbreviation...)

---

Table 7: BRIGHT-PRO Earth Science Example.



---

**Query:**

It's easy to understand why, as the Ukraine war disrupts global oil markets, gasoline prices are rising. What puzzles me is that E85, which is at most 49% gasoline, rises the same amount. It seems that it would be affected at most half as much as gasoline. Do we import a lot of ethanol from Russia?

I'm curious to know if there's a valid economic basis for this that doesn't involve price gouging.

---

**Reasoning Aspect 1 (weight = 0.18)**

The price connection is a domestic market phenomenon, as the U.S. is a net exporter of ethanol and does not import it from Russia.

**Positive Documents**

**Doc 1:** (...abbreviation...) The United States produces most of the fuel ethanol that it consumes. Production has consistently exceeded consumption since 2010, making the U.S. a net exporter of ethanol. Imports, mainly from Brazil, are minor and concentrated on the West Coast. (...abbreviation...)

---

**Reasoning Aspect 2 (weight = 0.36)**

Retailers price E85 at a consistent discount to regular gasoline to compensate drivers for its lower energy content and to compete on a cost-per-mile basis.

**Positive Documents**

**Doc 1:** (...abbreviation...) Ethanol has a lower energy content than gasoline, meaning that consumers get fewer miles per gallon when using E85. Retailers therefore maintain a price discount so that cost-per-mile is comparable to regular gasoline. (...abbreviation...)

**Doc 2:** (...abbreviation...) Surveys show that flex-fuel drivers generally switch to E85 only when its price is between 53% and 63% of E10's price, indicating that consumers require a substantial discount to offset reduced fuel economy. (...abbreviation...)

**Doc 3:** (...abbreviation...) Historical data show that E85 prices tend to fluctuate in proportion to gasoline prices because retailers preserve this constant energy-based discount margin, maintaining competitiveness. (...abbreviation...)

---

**Reasoning Aspect 3 (weight = 0.46)**

Wholesale economics and government policies—particularly the Renewable Fuel Standard (RFS) and its Renewable Identification Number (RIN) credit system—link the ethanol and gasoline markets, causing E85 prices to move with gasoline.

**Positive Documents**

**Doc 1:** (...abbreviation...) The pricing of E85 relative to gasoline depends on both ethanol production costs (driven by corn prices) and crude oil prices (which drive gasoline prices). The RFS program adds a coupling mechanism through RIN credits, which make E85 relatively cheaper when RIN values rise. (...abbreviation...)

**Doc 2:** (...abbreviation...) A linear model of E85 prices shows that  $E85 = 0.75 + 0.74 \times \text{ethanol price} + 0.26 \times \text{gasoline price} - 0.71 \times \text{RIN price}$ , implying strong dependence on both ethanol and gasoline prices and partial offset by RIN credits. (...abbreviation...)

**Doc 3:** (...abbreviation...) The RIN system acts as a tax on fuels with high petroleum content and a subsidy on high-ethanol fuels, effectively transmitting changes in gasoline and ethanol prices through both wholesale and retail levels. (...abbreviation...)

---

Table 8: BRIGHT-PRO Economics Example.

---

**Query:** What is the mechanism that makes people more likely to act when provided with proof of non-new (and often trivial) information?

---

**Reasoning Aspect 1 (weight = 0.36)**

External triggers or *cues to action* can prompt people to act on existing beliefs, moving them from belief to behavior without adding new information.

**Positive Documents**

**Doc 1:** (...abbreviation...) *Health Belief Model (HBM)*: core constructs (susceptibility, severity, benefits, barriers, self-efficacy) plus *cues to action* added in the 1970s to capture stimuli that initiate behavior. (...abbreviation...)

**Doc 2:** (...abbreviation...) *HBM details*: internal/external cues (e.g., noticing symptoms, exposure to a health campaign) trigger action even when beliefs already exist. (...abbreviation...)

**Doc 3:** (...abbreviation...) *Availability heuristic*: readily recalled consequences feel more important; making outcomes mentally available can nudge action. (...abbreviation...)

---

**Reasoning Aspect 2 (weight = 0.18)**

Fear-appeal models predict that messages pairing a serious threat with high perceived efficacy most reliably produce protective action.

**Positive Documents**

**Doc 1:** (...abbreviation...) *Extended Parallel Process Model (EPPM)*: high threat + high efficacy  $\Rightarrow$  danger-control responses (adaptive action); low efficacy risks fear-control responses. Meta-analyses show strongest effects when efficacy is high. (...abbreviation...)

---

**Reasoning Aspect 3 (weight = 0.18)**

Graphic warnings work largely by evoking negative affect and capturing attention—motivating behavior via affect-based processing rather than changing explicit beliefs.

**Positive Documents**

**Doc 1:** (...abbreviation...) *Affect in decision making*: fast, experiential, emotion-driven processing often guides choices ahead of deliberation (Zajonc; dual-process accounts; somatic markers). (...abbreviation...)

**Doc 2:** (...abbreviation...) Large U.S. trial: pictorial warnings increased quit attempts mainly through aversive reactions and salience/recall, with little effect on risk beliefs. (...abbreviation...)

**Doc 3:** (...abbreviation...) Meta-analysis: pictorial > text warnings on attention, negative affect, cognitive elaboration, quit intentions, and perceived effectiveness. (...abbreviation...)

**Doc 4:** (...abbreviation...) Experimental study: graphic (image+text) packs shift *implicit* smoking attitudes more than text-only, aligning with dual-process models. (...abbreviation...)

---

**Reasoning Aspect 4 (weight = 0.28)**

Making risks salient at the right moment (reminders) overcomes inattention and brings long-term goals to the forefront, shifting immediate decisions.

**Positive Documents**

**Doc 1:** (...abbreviation...) *Reminder/salience model*: reminders increase attention to future consequences and planned actions; when attention is limited, reminders raise saving and goal-consistent behavior. (...abbreviation...)

---

Table 9: BRIGHT-PRO Psychology Example.

---

**Query:**

Running ros1 packages with ros2 data: “I’m trying to run different lidar based SLAM methods to produce a map. My data are in .db3 and .mcap, but many SLAM methods (e.g., SMRT-AIST Interactive SLAM) are ROS1-only and expect .bag. Can I convert modern formats to .bag, or use older ROS1 packages with ROS2 datatypes?”

---

**Reasoning Aspect 1 (weight = 0.5)**

For data in the .mcap format, a two-step path is common: convert MCAP → ROS 2 .db3 with `ros2 bag convert`, then convert ROS 2 .db3 → ROS 1 .bag (using a bag conversion tool).

**Positive Documents**

**Doc 1:** (... abbreviation ...) Rosbag2 “convert” tool — `ros2 bag convert` can rewrite bags, merge/split, change compression, and perform serialization format conversion; example usage with multiple inputs and YAML `output_options`. (... abbreviation ...)

**Doc 2:** (... abbreviation ...) MCAP CLI — `mcap convert demo.db3 demo.mcap`; also supports converting ROS 1 .bag and ROS 2 .db3 to MCAP and notes schema caveats for pre-Iron .db3; mentions that `ros2 bag convert` can convert between .db3 and .mcap. (... abbreviation ...)

**Doc 3:** (... abbreviation ...) *rosbags (Python library)* — provides ROS1/ROS2 readers/writers and an efficient converter between rosbag1 and rosbag2; `rosbags-convert` supports converting between .bag and rosbag2 layouts. (... abbreviation ...)

---

**Reasoning Aspect 2 (weight = 0.5)**

Instead of offline conversion, you can play the ROS 2 bag and use `rosl_bridge` so ROS 2 topics appear live to ROS 1 nodes; then record or consume them with ROS 1 tools.

**Positive Documents**

**Doc 1:** (... abbreviation ...) *rosl\_bridge on Ubuntu 22.04* — guidance for building & using `rosl_bridge` to bridge ROS 2 releases with upstream ROS 1; outlines build-from-source steps and environment setup. (... abbreviation ...)

**Doc 2:** (... abbreviation ...) *Recording/playing with rosbag via bridge* — run ROS 1 `roscore`, start `rosl_bridge dynamic_bridge` with `-bridge-all-topics`, run ROS 2 publishers, then record bridged topics with `rosbag record`. (... abbreviation ...)

---

Table 10: BRIGHT-PRO Robotics Example.

<p><b>Query:</b> I have a custom PyTorch model that bottlenecks my application due to how it is currently used. The application is a web server built in Flask that receives job submissions for the PyTorch model to process. Due to the processing time of each job, I use Celery to handle the computation, where Flask queues the tasks for Celery to execute. Each job consists of loading the PyTorch model from the disk, moving the model and data to a GPU, and making a prediction on the data submitted. However, loading the model takes around 6 seconds. In many instances, that is a magnitude or two larger than prediction time.</p> <p>Thus, is it possible to load the model and move it to a GPU on server startup (specifically when the Celery worker starts), avoiding the time needed to load the model and copy it to the GPU every job? Ideally, I'd want to load the model and copy it to every available GPU on server startup, leaving each Celery job to choose an available GPU and copy the data over. Currently, I only have one GPU, so a multi-GPU solution is not a requirement at the moment, but I'm planning ahead.</p> <p>Further, the memory constraints of the model and data allow for only one job per GPU at a time, so I have a single Celery worker that processes jobs sequentially. This could reduce the complexity of the solution due to avoiding multiple jobs attempting to use the model in shared memory at the same time, so I figured I'd mention it.</p>
<p><b>Reasoning Aspect 1 (weight = 0.46)</b></p> <p>Use Celery startup/worker signals to load the model once per worker process (on the target GPU) so tasks reuse the in-memory model instead of reloading each time.</p> <p><b>Positive Documents</b></p> <p><b>Doc 1:</b> (...abbreviation...) Celery provides lifecycle signals like <code>worker_process_init</code>, <code>worker_ready</code>, and <code>celeryd_init</code>; initialize long-lived objects (e.g., models) when the worker process starts to avoid per-task setup. (...abbreviation...)</p>
<p><b>Reasoning Aspect 2 (weight = 0.18)</b></p> <p>Configure each Celery worker to a single, dedicated GPU and run with concurrency set to one to keep resource management simple and deterministic.</p> <p><b>Positive Documents</b></p> <p><b>Doc 1:</b> (...abbreviation...) PyTorch device-agnostic patterns and <code>CUDA_VISIBLE_DEVICES</code> allow pinning a worker to a specific GPU; use context managers or env vars to control device placement. (...abbreviation...)</p> <p><b>Doc 2:</b> (...abbreviation...) Practical use of <code>CUDA_VISIBLE_DEVICES</code> from shell/CLI (and notes about setting it before imports) to select GPU indices and avoid accidental multi-GPU access. (...abbreviation...)</p>
<p><b>Reasoning Aspect 3 (weight = 0.18)</b></p> <p>Optimize inference by using <code>torch.inference_mode()</code>, automatic mixed precision where safe, and non-blocking host→device transfers (optionally with pinned memory).</p> <p><b>Positive Documents</b></p> <p><b>Doc 1:</b> (...abbreviation...) <code>torch.inference_mode</code> disables gradient tracking and view/version bookkeeping for faster eval when no grads are needed. (...abbreviation...)</p> <p><b>Doc 2:</b> (...abbreviation...) <code>torch.amp.autocast</code> for mixed precision (FP16/bfloat16) can accelerate common layers during inference/training when accuracy permits. (...abbreviation...)</p> <p><b>Doc 3:</b> (...abbreviation...) Data transfer tips: <code>non_blocking=True</code> with appropriate streams; pinned memory can help but should be used judiciously; profile to validate speedups. (...abbreviation...)</p>
<p><b>Reasoning Aspect 4 (weight = 0.09)</b></p> <p>Manage worker lifecycle: recycle to mitigate leaks and enforce limits; alternatively, delegate serving to a dedicated model server when appropriate.</p> <p><b>Positive Documents</b></p> <p><b>Doc 1:</b> (...abbreviation...) Celery options like <code>-time-limit</code>, <code>-soft-time-limit</code>, <code>-max-tasks-per-child</code>, and <code>-max-memory-per-child</code> help contain runaway tasks and memory growth. (...abbreviation...)</p> <p><b>Doc 2:</b> (...abbreviation...) TorchServe notes for large/parallelized models; GPU assignment via <code>CUDA_VISIBLE_DEVICES</code> and parallel modes (PP/TP) for advanced serving scenarios. (...abbreviation...)</p>
<p><b>Reasoning Aspect 5 (weight = 0.09)</b></p> <p>For multi-GPU scaling, run one Celery worker per GPU and route tasks to per-GPU queues to simplify isolation and scheduling.</p> <p><b>Positive Documents</b></p> <p><b>Doc 1:</b> (...abbreviation...) Celery task routing/queues: create per-device queues and start workers with <code>-Q</code> to consume only their assigned queue; supports simple horizontal scaling. (...abbreviation...)</p>

Table 11: BRIGHT-PRO Stackoverflow Example.



---

**Query:** Perennial crop purely for compost for a no dig bed - what crop and what area?

---

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

A mix of perennial crops like Russian comfrey for “greens” and miscanthus or willow for “browns” can provide a sustainable source for compost.

**Positive Documents**

**Doc 1:** (...abbreviation...) Comfrey (*Symphytum spp.*) is easy to grow; Russian comfrey (*S. x uplandicum*) is recommended for organic fertiliser; small beds can be cut several times per year. (...abbreviation...)

**Doc 2:** (...abbreviation...) *Miscanthus × giganteus* is a tall, perennial grass with high dry-matter yields and environmental benefits (reduced runoff, erosion). (...abbreviation...)

**Doc 3:** (...abbreviation...) Short-rotation coppice (SRC) systems: willow, poplar, and broadleaf coppice species suited to periodic cutting and biomass supply. (...abbreviation...)

**Doc 4:** (...abbreviation...) Crop information for SRC willow (*Salix spp.*): establishment, management, cutting cycles, and long-term productivity for biomass. (...abbreviation...)

---

**Reasoning Aspect 2 (weight = 0.41)**

Specific perennial crops like comfrey, clover-ryegrass leys, and miscanthus have documented annual dry matter yields useful for planning growing area to meet a compost volume target.

**Positive Documents**

**Doc 1:** (...abbreviation...) Comfrey ‘Bocking 14’ 2016 records: four cuts totaling ~2.31 kg dry matter per plant per season (site example), indicating strong per-plant output. (...abbreviation...) (*perennial\_crop/comfreytrialshtml\_3.txt*)

**Doc 2:** (...abbreviation...) Grass-clover ley (organic) field trials (Denmark, 2017–2019): 5.8–11.8 Mg ha<sup>-1</sup> DM with N-fertilization gradients; clover proportion affects yield and protein. (...abbreviation...)

**Doc 3:** (...abbreviation...) Miscanthus performance: high yield potential on marginal land; management and fertilization (incl. sewage sludge) explored for yield optimization. (...abbreviation...)

**Doc 4:** (...abbreviation...) Miscanthus spring-harvest study: dry matter yields around ~20.5 t ha<sup>-1</sup> (treatment means), non-significant differences across sludge dosages in one trial. (...abbreviation...)

**Doc 5:** (...abbreviation...) Comfrey agronomic/nutrition data: reports of 0.78 kg m<sup>-2</sup> DM in one cut; literature ranges for yields and composition across *Symphytum spp.* (...abbreviation...)

**Doc 6:** (...abbreviation...) Minnesota trials summary: comfrey annual dry-weight yields ~3–12 t acre<sup>-1</sup> (multiple cuts), moisture ~87–92%; comparative yields vs. alfalfa/other forages. (...abbreviation...)

---

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

The ideal compost pile starts at C:N ≈ 30–40:1, achieved by mixing more browns than greens by volume; comfrey’s low C:N means it’s best as an activator, not the sole feedstock.

**Positive Documents**

**Doc 1:** (...abbreviation...) Composting fundamentals: maintain oxygen 5–15%, moisture 50–60%, and C:N 30–40:1 for efficient microbial activity and reduced odor/pathogens. (...abbreviation...)

**Doc 2:** (...abbreviation...) Backyard composting guidance: mix 2–3 parts browns to 1 part greens by volume; cover greens with 4–8 inches of browns; turn and manage moisture/air. (...abbreviation...)

**Doc 3:** (...abbreviation...) Comfrey nutrient profile: low C:N (~9.8:1) and high K; excellent activator and liquid feed; not ideal as sole compost material due to low C:N. (...abbreviation...)

---

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

The required growing area depends on finished-compost properties (bulk density, moisture) which determine how much raw biomass is needed to produce a target volume (e.g., 1 m<sup>3</sup>/yr).

**Positive Documents**

**Doc 1:** (...abbreviation...) Compost bulk density typically 500–700 kg m<sup>-3</sup> at 35–55% moisture; used to convert between dry-weight needs and volume targets for application/planning. (...abbreviation...)

---

Table 12: **BRIGHT-PRO** Sustainable Living Example.

## E Experiment Results

Model	BRIGHT Overall	Biology	Earth Science	Economics	Psychology	Robotics	Stack Overflow	Sustainable Living	Overall
DIVER-4B	<b>28.9</b>	<b>88.6</b>	<b>91.8</b>	<b>77.5</b>	<b>82.7</b>	<b>74.1</b>	<b>84.6</b>	<b>79.8</b>	<b>82.7</b>
Inst-XL	18.9	69.4	<u>80.3</u>	<u>65.4</u>	<u>72.1</u>	<u>69.0</u>	67.9	69.0	<u>70.4</u>
GTE-7B	22.5	<u>80.7</u>	79.8	63.7	66.3	60.4	62.4	59.5	67.6
Qwen3-8B	23.7	59.5	74.7	61.1	64.3	67.8	66.4	60.7	64.9
OpenAI-Embed-Large	17.9	68.0	75.8	63.3	70.2	61.8	67.3	66.4	67.5
GritLM	21.0	69.9	76.5	61.4	60.6	61.2	60.7	60.4	64.4
BM25	14.5	63.4	70.5	63.5	50.7	63.2	62.5	65.8	62.8
ReasonIR-8B	<u>24.4</u>	62.3	71.4	56.2	54.2	63.4	69.4	59.1	62.3

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
Diver	<b>28.9</b>	<b>73.2</b>	<b>76.5</b>	<b>56.8</b>	<b>64.1</b>	<b>55.9</b>	<b>74.4</b>	<b>52.7</b>	<b>64.8</b>
Inst-XL	18.9	48.9	58.8	<u>46.1</u>	<u>49.4</u>	47.7	58.8	<u>41.1</u>	<u>50.1</u>
GTE-7B	22.5	<u>61.8</u>	<u>64.9</u>	41.3	44.6	38.2	52.1	36.6	48.5
Qwen3-8B	23.7	38.8	57.4	43.9	46.3	<u>50.1</u>	56.9	36.5	47.1
Openai-Embed-3L	17.9	48.0	54.8	43.4	48.8	40.4	53.3	39.5	46.9
GritLM	21.0	51.2	58.5	41.1	43.7	43.9	49.5	35.5	46.2
BM25	14.5	42.1	49.4	41.6	32.3	43.5	50.9	39.9	42.8
ReasonIR-8B	<u>24.4</u>	42.5	51.9	37.9	39.4	43.9	<u>59.3</u>	34.2	44.2

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
Diver	<b>28.9</b>	<b>66.4</b>	<b>70.6</b>	<b>51.3</b>	<b>57.3</b>	<b>51.6</b>	<b>61.2</b>	<b>50.1</b>	<b>58.3</b>
Inst-XL	18.9	42.3	51.7	<u>42.1</u>	<u>42.2</u>	41.7	47.9	37.2	<u>43.5</u>
GTE-7B	22.5	<u>57.0</u>	<u>60.1</u>	35.0	35.4	37.8	39.9	32.8	42.6
Qwen3-8B	23.7	35.6	53.3	38.8	40.2	<u>45.5</u>	<u>48.1</u>	33.6	42.2
Openai-Embed-3L	17.9	43.2	49.1	37.0	38.7	<u>36.5</u>	<u>43.1</u>	37.2	40.7
GritLM	21.0	47.2	55.0	37.1	35.1	40.2	37.8	33.7	40.9
BM25	14.5	39.1	45.6	37.4	28.5	38.0	40.2	<u>37.2</u>	38.0
ReasonIR-8B	<u>24.4</u>	38.0	48.0	32.3	32.4	38.4	45.1	30.9	37.9

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

### Case Study: Why did Federal Reserve balance-sheet *capital* drop by 32% in Dec 2015?

Aspect coverage			Gold documents found by <i>Diver</i>
Aspect	Diver retrieved docs	ReasonIR retrieved docs	
<b>Post-2015 stability:</b> Capital levels have stayed relatively flat because the surplus is hard-capped by statute and the other major component—paid-in capital from member banks—changes very slowly.	2 / 3	0 / 3	<ul style="list-style-type: none"> <li>economics-41/extraction_1.txt, economics-41/extraction_2.txt: FAST Act + Treasury transfer.</li> <li>economics-41/extraction_5.txt, economics-41/extraction_6.txt: Surplus cap and slow-moving paid-in capital.</li> </ul>
<b>Dec 2015 driver (FAST Act):</b> The Fixing America's Surface Transportation (FAST) Act capped the Federal Reserve's surplus and required a large transfer to the U.S. Treasury (\$19.3B).	2 / 4	0 / 4	<ul style="list-style-type: none"> <li>economics-41/extraction_8.txt: "Deferred asset" accounting practice.</li> </ul>
<b>Accounting stabilization:</b> When net income is negative, the Fed books a deferred asset rather than reducing capital, further stabilizing the capital account.	1 / 2	0 / 2	

#### Model Responses (excerpts)

##### Diver Response (High Quality)

*"The 32% drop in the Fed 'capital' series in December 2015 was not a market loss or a sudden operating failure—it was a statutory remittance to the U.S. Treasury. Congress's FAST Act (enacted Dec. 4, 2015) capped aggregate Reserve Bank surplus at \$10 billion and required any excess be transferred... After the FAST Act remittance, the surplus account is effectively held near the statutory cap, and the other main component—paid-in capital—changes only slowly... In short, the cap makes the reported surplus much less variable..."*

##### ReasonIR Response (Low Quality)

*"Short answer— that large December 2015 fall in the Fed's 'capital' line on the FRED series (WCTCL) is almost certainly not an economic shock... but a book-keeping/data issue: either a one-time reclassification or data revision in the Fed's published Statement of Condition (H.4.1)... Because neither paid-in capital nor retained surplus has had large persistent shocks since then, the series has remained roughly flat after the 2015 change..."*

#### What the evidence shows

Diver successfully identified the legislative cause of the 2015 capital drop—the FAST Act's mandated transfer to the U.S. Treasury—and linked it to the observed data series in the Federal Reserve's balance sheet. Its retrieval covered all three relevant aspects: the one-time event (legislation and transfer), the subsequent statutory cap, and the structural stability of paid-in capital. As a result, its generated explanation precisely matched the true causal mechanism and quantitatively referenced the \$19.3B remittance.

In contrast, ReasonIR retrieved no gold documents and lacked exposure to the FAST Act evidence. Its response therefore defaulted to generic financial reasoning, attributing the drop to "data reclassification" or "bookkeeping issues." This reasoning pattern demonstrates a common failure mode when retrieval omits key causal sources—models generate plausible but incorrect narratives based solely on surface regularities.

Figure 6: Diver vs. ReasonIR on the query "Why did Federal Reserve capital drop 32% in Dec 2015, and why has it changed little since?"

## Case Study: Estimating the Probability of Contrail Formation from Flight and Weather Data

### Aspect coverage

Aspect	Gold docs retrieved
<b>Schmidt–Appleman criterion (SAC):</b> Provides deterministic equations for computing temperature and humidity thresholds required for contrail formation and persistence.	4 / 5
<b>Probabilistic modeling:</b> Deterministic thresholds are converted into probabilities via a logistic model using deviations from SAC thresholds and forecast uncertainties.	1 / 1
<b>Implementation details:</b> Incorporates vapor pressure formulas, aircraft propulsion efficiency, and microphysical corrections for accurate estimation.	1 / 2

### Gold documents retrieved

- earth-47/extraction\_0.txt  
earth-47/extraction\_1.txt  
earth-47/extraction\_2.txt  
contrail\_formation/gmdl.txt:  
Core SAC equations and threshold derivations.
- earth-47/extraction\_6.txt:  
Saturation vapor pressure formula (Murphy–Koop).
- earth-47/extraction\_4.txt:  
Engine efficiency and contrail factor modeling.

**Total Retrieve rounds: 3**

### Model Response (excerpt)

*“Yes — there is a well-established physical test, the Schmidt–Appleman criterion (SAC), which provides a deterministic threshold for whether aircraft exhaust and ambient air mixing will reach saturation and form visible contrails. To convert this into a probability, the deterministic SAC can be combined with a logistic model that accounts for deviations between the ambient temperature  $T$  and the critical temperature  $T_c$ , relative humidity with respect to ice (RH<sub>i</sub>), propulsion efficiency, and dynamical factors such as vertical motion and shear. A practical formulation is:*

$$P(\text{contrail}) = \frac{1}{1 + e^{-z}}, \quad z = a_0 + a_1(T_c - T) + a_2(\text{RH}_i - 1) + a_3\eta + a_4w + a_5\text{shear} + \dots$$

*Here,  $(T_c - T)$  and  $(\text{RH}_i - 1)$  are dominant predictors, while aircraft efficiency  $\eta$  and vertical dynamics modulate persistence. Implementation requires computing  $T_c$  from SAC using accurate saturation vapor pressure formulations (e.g., Murphy–Koop). The probabilistic model improves robustness under forecast uncertainty.”*

### What the evidence shows

The model comprehensively applied the Schmidt–Appleman criterion and retrieved all relevant gold documents covering the physical derivation, humidity dependence, and aircraft efficiency factors. Its explanation combined deterministic thermodynamic reasoning with a probabilistic framing, showing how logistic regression transforms the binary SAC test into a continuous probability of contrail formation.

By referencing the correct vapor pressure formula, aircraft propulsion efficiency, and dynamic meteorological influences, the response demonstrated a thorough, physics-grounded synthesis aligned with scientific literature. It exemplifies robust reasoning where retrieval provided both mechanistic and implementation-level details, enabling an accurate and operationally useful answer.

Figure 7: Model response based on Diver retriever on estimating contrail formation probability from flight and weather data. The model retrieved most of the key gold documents, accurately integrating physical equations and probabilistic methods.



Case Study: What creates the feeling of ‘excess’ blood pressure to an area of the body?

**Repeated off-target retrievals (by document)**

Document ID	Type	Times
21704heart_33_0.txt	Non-gold / off-topic	7
Ghrelin_6_1.txt	Non-gold / off-topic	5
Baroreceptor_3.txt	Non-gold / off-topic	5
<b>Subtotal (non-gold)</b>		<b>17</b>

Document ID	Type	Times
Baroreceptor_0.txt	tangential/low-weight	12
<b>Subtotal (overused tangential)</b>		<b>12</b>

**What the evidence shows**

A weak retriever exhibits a *repetition bias*: once it latches onto a small set of sources, it keeps resurfacing them across rounds—regardless of whether they are off-topic, low-utility golds, or only partially relevant—because it fails to discover *new* evidence. This novelty deficit creates an exploration stall: the same anchors (e.g., . . . Baroreceptor\_0.txt) crowd out alternative cues that would broaden the evidence base, so the LLM has little fresh material to update its hypothesis. In other words, repetition is not merely a precision problem; it is a search-dynamics problem where feedback from earlier rounds narrows rather than expands the candidate set. The measured concentration of repeats makes this lock-in visible and explains why answer quality degrades even when some relevant evidence is present elsewhere in the corpus.

Figure 8: The figure quantifies repeated off-topic retrieval from ReasonIR on the query “What creates the feeling of ‘excess’ blood pressure to an area of the body?”