

DeepResearchGym: A Free, Transparent, and Reproducible Sandbox for Deep Research

João Coelho*
jmcoelho@andrew.cmu.edu
INESC-ID, Carnegie Mellon University
Pittsburgh, PA, USA

Jingjie Ning*
jening@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Jingyuan He
jingyuanh@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Kangrui Mao
karrym@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Abhijay Paladugu
apaladug@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Pranav Setlur
psetlur@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Jiahe Jin
jjiahe@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Jamie Callan
callan@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

João Magalhães
jm.magalhaes@fct.unl.pt
NOVA LINCS
Lisbon, Portugal

Bruno Martins
bruno.g.martins@tecnico.ulisboa.pt
INESC-ID, Instituto Superior Técnico
Lisbon, Portugal

Chenyan Xiong
cx@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

Deep research systems represent an emerging class of agentic information retrieval methods that generate comprehensive and well-supported reports to complex queries, and/or answers to hard-to-locate factual questions. However, most existing systems rely on dynamic commercial search APIs, which pose reproducibility and transparency challenges, in addition to high costs. To address these limitations, we introduce DEEPRESEARCHGYM as a free and open-source search sandbox for reproducible research on deep research systems. The sandbox provides a search API that indexes large-scale public web corpora, namely ClueWeb22 and FineWeb, using a state-of-the-art dense retriever and approximate nearest neighbor search via DiskANN. It achieves comparable latency to popular commercial APIs while ensuring stable document rankings across runs. We demonstrate the sandbox’s utility through two use cases. For training, we synthesize queries grounded in the indexed corpora and show that search agents trained within the sandbox generalize to commercial search at inference time, enabling cost-effective reinforcement learning. For evaluation, we extend the Researchy Questions benchmark with LLM-as-a-judge metrics to measure alignment with users’ information needs, retrieval faithfulness, and report quality. Evaluation results show that system rankings remain consistent when switching from commercial APIs to ours.

*These authors contributed equally to this work.



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICTIR '26, Melbourne, VIC, Australia
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2600-2/2026/07
<https://doi.org/10.1145/3805713.3820409>

CCS Concepts

• Information systems → Web searching and information discovery.

Keywords

Information Retrieval, Deep Research, Web Search

ACM Reference Format:

João Coelho, Jingjie Ning, Jingyuan He, Kangrui Mao, Abhijay Paladugu, Pranav Setlur, Jiahe Jin, Jamie Callan, João Magalhães, Bruno Martins, and Chenyan Xiong. 2026. DeepResearchGym: A Free, Transparent, and Reproducible Sandbox for Deep Research. In *Proceedings of the 2026 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '26)*, July 25, 2026, Melbourne, VIC, Australia. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3805713.3820409>

1 Introduction

Recent advances in Large Language Models (LLMs) have driven a transformation in information access paradigms, moving beyond ranked retrieval toward systems capable of synthesizing comprehensive report-style responses to complex queries. These deep research systems aim to address complex and open-ended information needs, combining iterative retrieval with multi-step reasoning and generation, and autonomously navigating and evaluating diverse sources to either construct well-supported reports or to find answers to hard-to-locate factoids. Prominent commercial examples include OpenAI [31] and Perplexity [35] deep research modes, which have demonstrated how these systems can significantly enhance user experience when addressing intricate questions requiring synthesis across multiple sources. Recent industry developments further underscore this shift, with both Google and Microsoft moving towards AI-driven search tools [6, 49].

As deep research systems become more common, they also introduce novel challenges for research. Being agentic by design, these systems rely on iterative search, retrieval, and reasoning over vast collections of online data, making both evaluation and training dependent on access to environments with diverse coverage, that faithfully simulate real-world behavior. Yet, such infrastructures remain scarce to the research community, forcing reliance on commercial web search APIs. While convenient, these APIs introduce critical limitations: their proprietary nature restricts transparency in the retrieval processes, their continuous evolution undermines reproducibility, and their cost prohibits large-scale training.

To address these challenges, we introduce DEEPRESEARCHGYM as a free and open-source search sandbox for reproducible research on deep research systems. At its core lies a search API built upon web snapshots comprising millions of documents, namely ClueWeb22 [33] and FineWeb [34]. The retrieval pipeline is composed entirely of publicly available components, namely the document collections, a state-of-the-art embedding model, and a scalable approximate nearest neighbor search index. This enables researchers to audit system behavior, analyze retrieved evidence, and rerun experiments under stable and reproducible search conditions. We also provide code to support local deployment, allowing experiments with different retrieval models and document collections. We highlight three core contributions:

- **Open search infrastructure.** We release an open search API that achieves strong retrieval quality with minimal loss from approximate search, while maintaining response times comparable to commercial APIs. Unlike volatile proprietary services, retrieval results remain stable over time, ensuring reproducibility across experimental runs.
- **Corpus-grounded query generation for training.** We synthesize queries grounded in the indexed corpora and use them to train search agents entirely within the sandbox. Training on the live web typically requires thousands of costly API calls and yields non-reproducible results. We show that agents trained within DEEPRESEARCHGYM generalize effectively to commercial search APIs at inference time, achieving learning gains comparable to agents trained directly on commercial systems.
- **Evaluation benchmark for deep research.** Building upon the Researchy Questions dataset [38], we adopt an LLM-as-a-judge framework to assess generated reports along multiple dimensions, including alignment with user intent, factual grounding, and overall report quality. Results show that systems using our API maintain report quality comparable to those relying on commercial search services.

Our findings support DEEPRESEARCHGYM as a practical and scalable sandbox for deep research, validated by an adoption of approximately 14.4 million API queries processed during its first six months of availability. Code, data, and documentation are available at <https://www.deeprsearchgym.ai>.

2 Related Work

Early work on Retrieval-Augmented Generation (RAG) systems focused on improving performance on knowledge-intensive question answering, by retrieving supporting documents from large

corpora and conditioning generation on this evidence, in order to enhance factual accuracy [21, 50, 58]. Building on this foundation, several deep research systems have been optimized for short-form factoid-style question answering. These include reinforcement learning approaches that enable search agents to autonomously navigate the web, issue iterative queries, and synthesize concise responses [17, 46, 57], as well as prompt-based methods like Search-o1 [23], which equip LLMs with the ability to trigger web searches when encountering knowledge gaps.

A complementary line of work focuses on long-form report generation frameworks. GPTResearcher [11] orchestrates agentic workflows to coordinate planning, retrieval, and drafting across hybrid data sources, incorporating techniques such as report planning [52] and query decomposition [8] to enhance long-form synthesis, while enforcing completeness. Building on these paradigms, other deep research systems emphasize agentic tool use to extend reasoning capabilities beyond pure text-based retrieval [14, 29]. For instance, OpenDeepSearch [2] combines iterative retrieval with code execution for complex tasks. HuggingFace’s OpenDeepResearch initiative [16] follows similar directions in an open-source framework, while emphasizing transparency and modularity.

A common limitation across the aforementioned systems is their reliance on commercial web search APIs, such as Tavily [48] or Serper [43], for document retrieval. These APIs provide limited transparency into document indexing and ranking, are subject to changes over time, and restrict the ability to replicate retrieval conditions, posing challenges for reproducibility and evaluation. Furthermore, several such services rely on large-scale scraping of proprietary search engine result pages, particularly Google SERPs, rather than maintaining independent search indexing and ranking pipelines. Recent legal action by Google against SERP scraping services [7] underscores the fragility of this dependency, motivating the need for research-oriented retrieval infrastructures. These limitations extend beyond inference to training, as recent reinforcement learning approaches for search agents require thousands of API calls per experiment [24, 46, 57], coupling both cost and reproducibility to commercial services.

Parallel efforts have targeted the evaluation of deep research systems’ quality. In particular, multiple benchmarks have driven progress on short-form question answering, such as GAIA [26], HLE [36], and FRAMES [20]. Concurrent to our work, BrowseCompPlus [10] curates a fixed 100K-document corpus for short-form factoid QA, aiming to isolate retriever contributions to answer correctness. Our framework differs in scope, providing a search API over large-scale web corpora that supports cost-effective agent training and report evaluation under reproducible retrieval conditions.

Recent work has also introduced frameworks that addressed the challenges of evaluating long-form synthesis. FACTScore [27] and SAFE [53] decompose outputs into atomic claims and verify their factual consistency against external sources. For retrieval-augmented systems, ARES [41] and RAGChecker [40] offer modular evaluations that explicitly link generated claims to retrieved evidence. Long²RAG [37] extends this approach by introducing Key Point Recall (KPR), which evaluates how well long-form answers capture essential content from retrieved sources by measuring coverage of salient points.

Table 1: Retrieval performance of DEEPRESEARCHGYM’s search API, measured over the Researchy Questions test set.

L	Relevance Eval			ANNS Eval	
	MRR@10	nDCG@10	R@100	R@10	R@100
100	48.34	39.40	78.06	90.01	88.72
200	48.39	39.49	78.27	92.63	91.01
300	48.41	39.50	78.35	93.87	92.64
400	48.44	39.52	78.39	94.72	93.68
500	48.45	39.55	78.43	95.39	94.39

3 DeepResearchGym: API Search Sandbox

This section presents DEEPRESEARCHGYM as an open-source framework designed to support reproducible research on deep research systems. To address the challenges related to the reliance on commercial web search APIs, DEEPRESEARCHGYM offers a controlled sandbox environment built on large-scale web corpora. We begin by describing the underlying web corpora, followed by an overview of the dense retriever and the ANN indexing approach used to enable efficient search. Finally, we outline the API interface, including available endpoints and supported arguments.

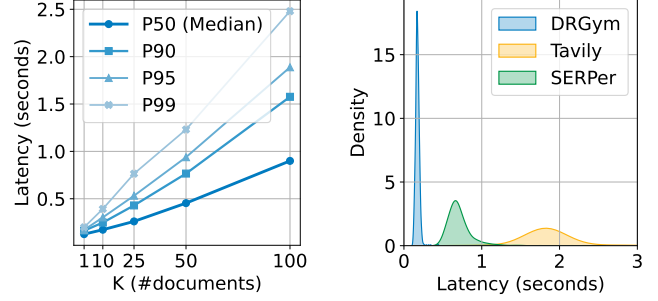
3.1 Web Corpora

DEEPRESEARCHGYM indexes three large-scale web datasets, namely the English subsets of ClueWeb22 A and B [33], and the FineWeb CC-MAIN-2024-51 snapshot [34].

ClueWeb22 was collected in 2022 and comprises approximately 10 billion web pages. It is organized into three categories, each representing different segments of the web. Category B, known as ClueWeb22-B, approximates the *super head* of the web, encompassing the most frequently visited pages (e.g., pages from Wikipedia, major news outlets, and other top domains). It includes around 200 million web pages, with approximately 87 million in English. These pages were sampled based on their likelihood to satisfy user information needs, as estimated by a commercial search engine’s importance scoring. Low-quality and spam pages were filtered during sampling to enhance the dataset’s overall quality.

To mitigate potential coverage concerns and ensure that systems can be exposed to a broader spectrum of web content, we also provide access to ClueWeb22-A. This larger subset encompasses approximately 1 billion English pages from the *mostly head* of the web, offering a more diverse mix of frequently visited websites.

In turn, FineWeb is a large-scale English web corpus collected from 96 Common Crawl snapshots between 2013 and 2024. It comprises approximately 15 trillion tokens of cleaned and deduplicated web data. The dataset employs rigorous filtering, deduplication, and quality control measures, resulting in a high-quality resource for tasks such as LLM training. To mitigate temporal constraints associated with ClueWeb22, we focus on the latest 2024 crawl, which includes over 180 million documents capturing more recent trends

**Figure 1: Latency percentiles while varying K for DEEPRESEARCHGYM (left), and latency comparison with two commercial APIs for $K = 10$ (right).**

compared to earlier data. This makes the collection valuable for queries that require up-to-date information, reflecting the evolving nature of web content and user interests.

3.2 Search Indexes

To enable efficient state-of-the-art retrieval across the above corpora, we built a dense retrieval backend combining state-of-the-art embedding models and approximate nearest neighbor search.

Our choice of embedding model was guided by three requirements: strong retrieval performance across diverse domains, support for long document contexts, and openly available model weights. Among the models we tested, MiniCPM-Embedding-Light [15, 32] best satisfied these criteria. The model leverages bidirectional attention mechanisms [4] and weighted mean pooling [28], supporting documents with up to 8192 tokens. On the BEIR benchmark [51], it achieves a zero-shot nDCG@10 of 55.27, outperforming alternatives such as bge-large-en-v1.5 [3] and jina-embeddings-v3 [19]. While we serve MiniCPM-Embedding-Light as the default model, all code for indexing and serving is open-source, allowing researchers to substitute with a retriever of their choice.

We index the document embeddings using DiskANN [47]. Each corpus is partitioned into independent shards of 25 million documents, which are separately indexed for distributed deployment. During search, shards are queried in parallel, and the top-ranked results are merged to produce a final ranking.

To assess the retrieval effectiveness of our search system, we evaluated it on the Researchy Questions test set. This experiment used ClueWeb22-B as the corpus, since the Researchy Questions relevance labels are grounded on it. Table 1 presents the retrieval performance, considering the number of retrieved documents $K = 100$, while varying L , i.e. a DiskANN search-time parameter that controls the size of the candidate neighbor list explored during search. Increasing L boosts ranking quality by allowing for more exploration of the search graph, but reduces query throughput. We provide metrics computed given the ground-truth clicked documents (MRR@n, nDCG@n, and R@n), as well as approximate nearest neighbor recall (ANN R@n), computed based on exact-search results. The results with increased L indicate that the error introduced by ANN search is low, thus solidifying the retrieval quality of our backend.

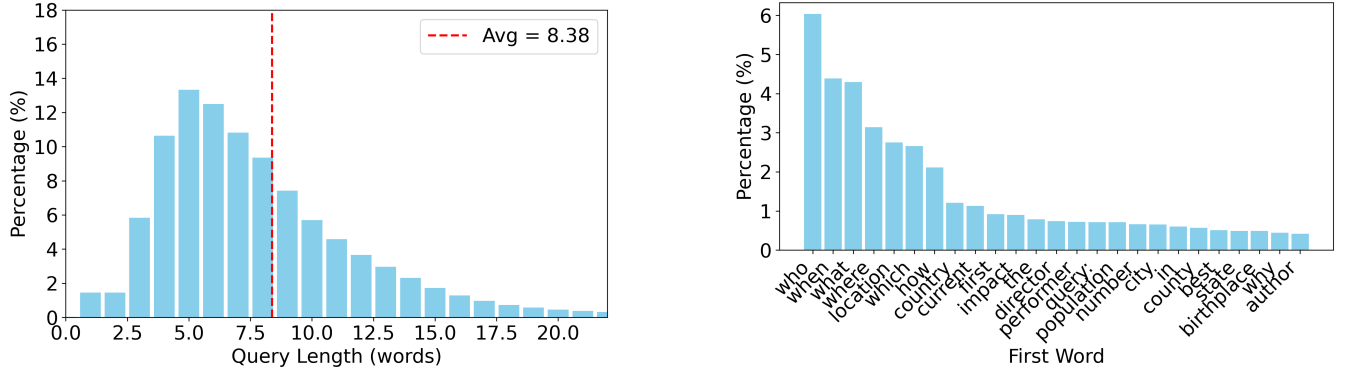


Figure 2: Query log analysis: length distribution (left), and 25 most common first-words (right).

3.3 Retrieval API

DEEPRESEARCHGYM provides a retrieval API designed to support deep research systems over the aforementioned corpora. The API exposes two primary endpoints: (i) a `/search` endpoint, which accepts a text query and returns a ranked list of documents from the selected corpus, and (ii) a `/fetch` endpoint, which retrieves the archived textual content of a document given its URL. Users can choose which corpus to use through an API parameter.

The `/search` endpoint supports document retrieval over the previously introduced corpora. As for search-time DiskANN parameters, our API defaults to a dynamic behavior of $L = K \times 5$, since, by definition, $\min(L) = K$. Since deep research systems typically issue queries sequentially rather than in batches, we evaluate our API’s latency in this single-query setting, comparing it to commercial alternatives. Figure 1 presents the results, with the left panel showing percentile-based end-to-end latency for our API across different values of K (the number of retrieved documents), and with the right panel comparing latency against commercial APIs for $K = 10$, i.e. a common setting for deep research systems. Our API consistently yields sub-second latency, outperforming commercial search services.

In turn, the `/fetch` endpoint addresses a specific challenge in deep research systems supported by static web corpora. During generation, systems retrieve documents via the `/search` endpoint, accessing versions captured during the crawl. Their final reports cite the original URLs associated with these documents. However, the live content of such URLs may have changed or disappeared since the original crawl. To mitigate this discrepancy, the `/fetch` endpoint serves archived snapshots of documents as captured during the crawl, ensuring that the original content of URLs cited in reports can be retrieved post-generation. This design enables the construction of isolated deep research pipelines that are independent of dynamic or degraded external sources. The endpoint maintains a median latency of 0.09 seconds per single request.

4 DeepResearchGym Usage Analysis

During its first six months of availability, our API recorded 14.4 million search requests from 558 unique IP addresses across 25

countries. We analyzed this query log to characterize usage patterns, presenting key statistics here. In separate work, Ning et al. [30] detail a large-scale analysis of the logs from DEEPRESEARCHGYM, studying how search sessions unfold and retrieved evidence is used.

We first classified a representative sample of queries in our log using a standard web search taxonomy that considers informational, navigational, and transactional intents. Prior studies of general-purpose search engines report that around 50% of queries are informational [9]. In contrast, our logs reveal that approximately 90% of the queries are informational, suggesting that users primarily employed the service for knowledge-seeking purposes, as expected from deep research systems.

Figure 2 (left) shows the query length distribution. The average query length is 8.38 words, which is considerably higher than that of typical web search queries, generally reported between 3 and 4 words [5, 39]. This further suggests that queries are more complex, consistent with the goals of deep research systems. Figure 2 (right) displays the top-25 most frequent first words, highlighting a dominance of WH-questions (e.g., who, when, and what) and geo-spatial terms (e.g., location, country), another pattern consistent with information-seeking behavior.

5 DeepResearchGym for Search Agent Training

Training search agents through reinforcement learning needs thousands of search API calls. For instance, an experiment with 10,000 queries, 16 trajectories per query, and 4 searches per trajectory, results in 640,000 API calls – costing up to US\$640 with Serper or US\$5,000 with Tavily. For multiple experimental runs this becomes expensive, besides being subject to evolving API behaviors. In contrast, DeepResearchGym is free and returns stable results across runs, enabling cost-effective and reproducible experimentation.

In this section, we empirically show that agents trained using DEEPRESEARCHGYM’s search API generalize to both unseen benchmarks and commercial search environments, validating the framework as a practical free training environment. We focus on short-form question answering rather than long-form report generation, as this setup is more efficient to train and rewards are easier to establish, while maintaining the core search behaviors.

Table 2: Search agent performance before and after reinforcement learning with different search engines. The entries marked with “-” correspond to cases where the benchmark requires live web content, unavailable in ClueWeb22.

Method	Search Train	Search Infer	GAIA	WebWalker	HLE	ClueWeb-Test
Qwen3-1.7B [18]	(no training)	Serper	8.7	19.1	6.2	16.9
Qwen3-1.7B [18]	(no training)	ClueWeb	-	-	-	13.2
Qwen3-1.7B + RL (Live Web)	Serper	Serper	18.4	35.4	6.9	24.5
Qwen3-1.7B + RL (DEEPRESEARCHGYM)	ClueWeb	ClueWeb	-	-	-	18.8
Qwen3-1.7B + RL (DEEPRESEARCHGYM)	ClueWeb	Serper	20.3	29.7	7.0	22.6

5.1 Sandboxed Query Synthesis

Existing approaches synthesize training queries grounded on the live web, either by starting from an arbitrary document and corresponding link structure, or by explicitly leveraging a commercial search engine during multi-hop query generation [12, 22, 45]. This couples query generation to commercial APIs and produces queries that may not be answerable using static corpora.

We adapt these approaches to be grounded over ClueWeb22 instead. Given a root Wikipedia document from ClueWeb22, we prompt gpt-4.1-mini to generate queries targeting information in that document. To introduce multi-hop reasoning, we leverage the document’s link structure (i.e. inlinks and outlinks, restricted to ClueWeb22) to identify related entities for constructing questions requiring synthesis across multiple documents. We also employ LLM-based selective obfuscation (e.g., replacing “PageRank” with “Google’s original ranking algorithm”) to remove directly searchable terms while preserving enough information to identify the answer.

All generated queries undergo validation via LLM judging to ensure they are (i) answerable using our search API, (ii) have a single correct answer, and (iii) require search rather than relying solely on parametric knowledge.

5.2 Agent Training and Inference

Following prior work [18], we train a search agent with Qwen3-1.7B [56] as the backbone LLM, equipped with three actions: search, summarize context, or answer. Training follows GRPO [44], using LLM-as-a-judge soft-match rewards between agent responses and ground-truth answers. We compare two configurations: First, we train with a commercial search API (Serper) on synthetic data from AFM-Web-Agent [22]. Second, we consider the aforementioned synthetic questions grounded over ClueWeb22 rather than the live web, and train with DEEPRESEARCHGYM’s API. Both configurations train on 5,000 queries. The RL training is conducted for 150 steps, with a batch size of 32 and a group size of 8 for GRPO.

At inference time, we evaluate under two scenarios. First, to isolate the effect of training infrastructure, both agents are evaluated on standard benchmarks, using Serper to query the live web regardless of which API they were trained with.

Second, to provide an in-distribution comparison, we evaluate each agent using its respective training API on ClueWeb-Test, i.e. a curated subset of the benchmarks that is covered by ClueWeb22. We construct this subset by filtering GAIA, WebWalkerQA, and HLE with a validation agent (Gemini-2.5-Flash), retaining the 53 queries it solves at pass@5 while issuing at least one search call.

Table 3: Sample of the Researchy Questions dataset with hyperlinks to labeled relevant documents.

Query	References
Is the COVID vaccine dangerous	URL1 , URL2 , URL3 , URL4 , URL5
Why is there a chip shortage	URL1 , URL2 , URL3 , URL4 , URL5
Can there be knowledge that is independent of culture?	URL1 , URL2 , URL3 , URL4 , URL5

5.3 Results

Table 2 presents pass@1 success rates on three standard agentic search benchmarks: GAIA [26], WebWalkerQA [54], and HLE [36]. ClueWeb-Test is a smaller 53-query subset from these benchmarks, answerable using ClueWeb22, for which we report pass@4. Although this subset is small, both setups exhibit comparable relative improvements, which supports the use of DEEPRESEARCHGYM as a stable environment for probing training dynamics.

Furthermore, the ClueWeb-trained agent generalizes effectively to commercial search at inference, matching or exceeding the Serper-trained baseline on GAIA and HLE, though trailing on WebWalkerQA. While an expected performance gap exists between search backends, these results confirm that DEEPRESEARCHGYM enables cost-effective and reproducible training, where learned search strategies transfer across both engines and evaluation distributions.

6 DeepResearchGym for Report Evaluation

We further demonstrate the sandbox’s utility by instantiating a benchmarking protocol for long-form report generation. We evaluate multiple deep research systems and show that rankings remain consistent when switching from commercial APIs to DEEPRESEARCHGYM’s. While the sandbox is agnostic to the specific evaluation task, we introduce this protocol to fill a gap in the evaluation landscape, and provide empirical observations using our API.

6.1 Researchy Questions

Evaluating deep research systems requires queries that naturally drive extensive information exploration and synthesis. The Researchy Questions dataset [38] was curated specifically to capture such queries. Rather than featuring simple factoid questions, the dataset consists of approximately 96,000 real-world information-seeking queries that led users to engage with multiple documents during search sessions, as measured by aggregated click distributions over the ClueWeb22 dataset.

Table 4: Comparison of deep research systems on the Researchy Questions test set using (i) each system’s original commercial search API, (ii) DEEPRESEARCHGYM’s ClueWeb22 API, and (iii) DEEPRESEARCHGYM’s FineWeb API. Scores are judged by gpt-4.1-mini-2025-04-14. The system marked with * is not tailored for long-report generation.

System	Relevance						Faithfulness						Quality					
	Commercial KPR	Commercial KPC	ClueWeb KPR	ClueWeb KPC	FineWeb KPR	FineWeb KPC	Commercial Prec	Commercial Rec	ClueWeb Prec	ClueWeb Rec	FineWeb Prec	FineWeb Rec	Commercial Clar	Commercial Ins	ClueWeb Clar	ClueWeb Ins	FineWeb Clar	FineWeb Ins
perplexity-sonar-deep	72.50	1.12	–	–	–	–	55.65	99.22	–	–	–	–	89.50	89.26	–	–	–	–
gpt4-search-preview	40.01	1.69	–	–	–	–	57.68	56.11	–	–	–	–	70.13	59.13	–	–	–	–
GPTResearcher	60.61	1.52	64.67	1.42	64.47	1.54	89.11	94.29	85.36	90.82	87.42	92.41	86.37	81.52	83.70	78.01	82.90	77.73
OpenDeepSearch	32.92	0.97	42.81	0.84	40.73	0.86	85.86	97.78	81.32	94.82	82.70	95.19	59.20	47.04	61.48	49.51	61.25	50.01
HuggingFace-DeepSearch	33.00	0.81	35.22	1.35	38.16	1.32	0.35	0.29	0.10	0.10	0.12	0.13	57.52	47.98	58.34	52.36	55.81	47.62
Search-R1*	5.52	0.81	4.95	0.80	5.45	0.96	0.00	0.00	0.00	0.00	0.00	0.00	9.48	11.87	9.07	11.18	9.70	12.10

The heavy engagement with diverse sources reflects the essential challenges deep research systems are designed to address: synthesizing information across multiple perspectives, reconciling conflicting evidence, and constructing comprehensive responses. While the dataset was originally introduced for evaluating retrieval performance, its properties make it a strong foundation for studying long-form generation grounded in multi-document evidence. In the next subsection, we describe the proposed evaluation methodology, which extends the use of Researchy Questions to benchmark report generation with deep research systems.

Analyzing the queries (Table 3) reveals a trade-off between the reproducibility provided by static web corpora and the limited ability of such corpora to support evaluations that depend on recent events. Questions such as “*why is there a chip shortage*” may require post-cutoff updates, while queries such as “*can there be knowledge that is independent of culture*” rely on conceptual knowledge that remains stable across time. To quantify this, we categorized test queries into temporally stable versus temporally evolving, leveraging gpt-4.1-mini. The large majority (81.6%) are temporally stable, with only 18.4% involving evolving topics. In both cases, the static corpus provides controlled access to all ground-truth evidence available up to the snapshot date.

6.2 Long-form Report Evaluation Metrics

Deep research systems that focus on providing report-like answers face multiple challenges inherent to long-form generation evaluation [55], where outputs must be assessed not only for linguistic fluency and informativeness, but also for factual grounding and content relevance. We propose a tri-faceted evaluation framework that assesses the alignment with user information needs, factual grounding, and overall quality of generated answers.

Report Relevance: As the primary metric for assessing user satisfaction, we evaluate how well the generated reports address the user’s underlying information needs. Given that Researchy Questions instances are derived from real-world web search sessions, we leverage the set of documents clicked by users as a proxy for ground-truth information targets. Following the Key Point Recall (KPR) methodology [37], we extract salient points from each ground-truth document using an LLM guided by structured prompts, capturing the core content users engaged with. We then assess each generated report for semantic inclusion of these key points, computing the KPR score as $\frac{1}{M} \sum_{j=1}^M c_j$, where M is the total number of key points

and c_j indicates whether key point j is supported by the report, as judged by an LLM.

To complement recall, we also compute Key Point Contradiction (KPC), which measures whether the report introduces statements that conflict with any key points. This score captures potential misinformation or misleading content, defined as $\frac{1}{M} \sum_{j=1}^M d_j$, where d_j is 1 if the report contradicts key point j , as judged by the same LLM used for the KPR metric. Together, these metrics provide a user-centered assessment of both coverage and factual consistency relative to real-world search intents.

Retrieval Faithfulness: Beyond relevance, we assess the factual grounding of generated reports, adapting the LLM-as-a-judge approach of the TREC-RAG evaluation process [50]. Our citation evaluation pipeline follows a three-stage process. First, factual claims are extracted from the report, along with any URLs referenced as support for each claim. Second, the content of each cited source is retrieved from the corresponding URL(s). Finally, an LLM is prompted to assess whether the cited source adequately supports the corresponding claim.

Given a report, we compute the primary metrics established by the TREC-RAG evaluation. Citation recall measures the proportion of factual claims that include at least one citation, i.e., $\frac{N_{\text{cited}}}{N_{\text{total}}}$, where N_{cited} represents the number of claims with citations and N_{total} represents the total number of claims. This metric quantifies how consistently the system grounds its assertions in external evidence.

In turn, citation precision evaluates the quality of citations for claims that include references. Each claim-citation pair receives a support score s_i , where full support (score = 1) means all key aspects of the claim are fully supported by the cited source; partial support (score = 0.5) means some aspects of the claim are supported, but the support is incomplete; and no support (score = 0) means the cited source does not substantively support the claim or is irrelevant. Citation precision is then computed as the average score across all cited claims, i.e., $\frac{1}{N_{\text{cited}}} \sum_{i=1}^{N_{\text{cited}}} s_i$.

Report Quality: To capture aspects of writing quality and analytical depth, we employ an LLM-as-a-Judge protocol [13], prompting a strong LLM to evaluate each answer along two key dimensions: clarity, reflecting logical coherence and linguistic fluency; and insightfulness, capturing analytical nuance and the depth of reasoning presented. These dimensions are commonly used in long-form generation evaluation [25, 42] and provide evidence of the presentation quality of the generated content.

6.3 Experimental Setup

To evaluate the current landscape of deep research systems, we conducted a systematic benchmarking study, following the evaluation protocol described in the previous section. We used a subset of the previously introduced Researchy Questions dataset, namely the top 1,000 queries from the test set, ranked by the number of documents clicked during the original search sessions. This ranking naturally favors queries that drive extensive exploration, aligning with the overall goals of report-generating deep research systems.

We benchmark commercial systems (perplexity-sonar-deepsearch and gpt4-search-preview) and open-source systems (GPTRearcher, OpenDeepSearch, and HuggingFace-DeepSearch). We also include Search-R1 as a lower-bound reference, as it focuses on short-form answers. All systems are evaluated in default configurations using gpt-4.1-mini-2025-04-14 as the judge. For open-source systems, we evaluate under two retrieval settings. We consider each system’s default commercial API, as well as DeepResearchGym’s API, showing independent results for the ClueWeb and FineWeb corpus. For the commercial systems, the search backend cannot be re-routed due to their proprietary integration, so we report only their native search configuration.

6.4 System-Level Evaluation

Table 4 presents results under two retrieval configurations: each system’s original commercial API, and DEEPRESEARCHGYM’s API. Despite Researchy Questions being grounded on ClueWeb22, we also provide results when searching over FineWeb.

The results show that systems maintain their relative performance rankings across both retrieval settings, confirming that DEEPRESEARCHGYM’s search API provides sufficient retrieval quality to support effective report generation.

We also observe consistent patterns in the relative difficulty of different evaluation dimensions. Even top-performing systems like perplexity-sonar-deepsearch and GPTRearcher achieve notably higher scores in report quality metrics (Clarity, Insight) compared to information coverage metrics (KPR), suggesting that linguistic fluency has outpaced comprehensive content synthesis. This pattern holds across both retrieval environments, indicating an intrinsic challenge in deep research that transcends retrieval infrastructure.

The evaluation also reveals a trade-off in commercial systems, which tend to produce strong narrative quality but sometimes with reduced citation precision. Manual inspection of the results shows two recurring issues: (1) citations are often used to support broad sections instead of specific claims, and (2) some cited URLs cannot be fully crawled. This points to a tension between narrative coherence and precise evidence anchoring in current designs. For these citation metrics, note that systems that do not present references receive Faithfulness scores of 0.

We also evaluated the experiment summarized in Table 4 with alternative LLM judges (Gemini-2.5-Pro and GPT-OSS-20B). While absolute scores vary across judges, per-query assessments show strong correlation on relevance ($\rho \approx 0.8$) and moderate correlation on faithfulness and quality ($\rho \approx 0.5$).

6.5 Query-level Analysis

To further investigate the consistency of system performance across individual queries, we conducted a fine-grained analysis comparing results obtained under each system’s original retrieval API, against those obtained with DEEPRESEARCHGYM’s API, focusing only on the systems geared towards long-report generation and explicit references. Figure 3 presents scatter plots of per-query scores across our three evaluation axes.

The analysis reveals distinct patterns across evaluation dimensions. For relevance (KPR), stronger systems exhibit moderate to high correlation, indicating that query-level retrieval effectiveness is largely preserved when transitioning to DEEPRESEARCHGYM’s corpus. However, mid-range queries show some score variability, suggesting that certain information needs are more sensitive to differences in retrieval infrastructure. In contrast, report quality metrics demonstrate lower per-query correlation, despite high absolute scores for top systems. This implies that while narrative fluency and coherence are robust to retrieval changes, they are not tightly coupled with individual query characteristics.

Retrieval faithfulness shows the lowest per-query correlation across systems, indicating that this dimension is sensitive to differences in retrieved evidence. Changes in the retrieved documents can shift not only how well claims are supported, but also the claims themselves, leading to variation in citation faithfulness scores across retrieval setups. While average scores across queries remain stable, with some individual queries yielding consistently high scores across both sources, the broader pattern lacks alignment, with most points scattered and no clear linear trend. This variability underscores the importance of using a standard retrieval API when benchmarking deep research systems, as it helps control for retrieval effects and ensures that observed differences stem from model behavior rather than different access to evidence.

6.6 Human Evaluation

To validate our automatic evaluation protocol and assess its alignment with human judgments, we conduct a systematic human evaluation study. The evaluation set comprises 210 queries with their corresponding generated reports. For each query, we present annotators with two reports generated by different systems, asking them to select the better one based on overall informativeness, coherence, and factual accuracy. Seven annotators (all also co-authors of this paper) participated in the study, with each query independently evaluated by three different individuals. To mitigate potential bias, the evaluation was conducted double-blind: annotators were unaware of which systems produced each report, and system assignments were randomized across queries. Report presentation order was also randomized to prevent positional bias, and ties were disallowed to enforce binary preferences.

To assess inter-annotator reliability, we compute pairwise Cohen’s κ scores across all annotated comparisons. The average pairwise κ is 0.87, indicating a high level of agreement and supporting the consistency of human judgments. Building on this, Table 5 presents Cohen’s κ values measuring agreement between LLM-based automatic judgments and human preferences across each evaluation dimension, while Figure 4 presents pairwise win rates,

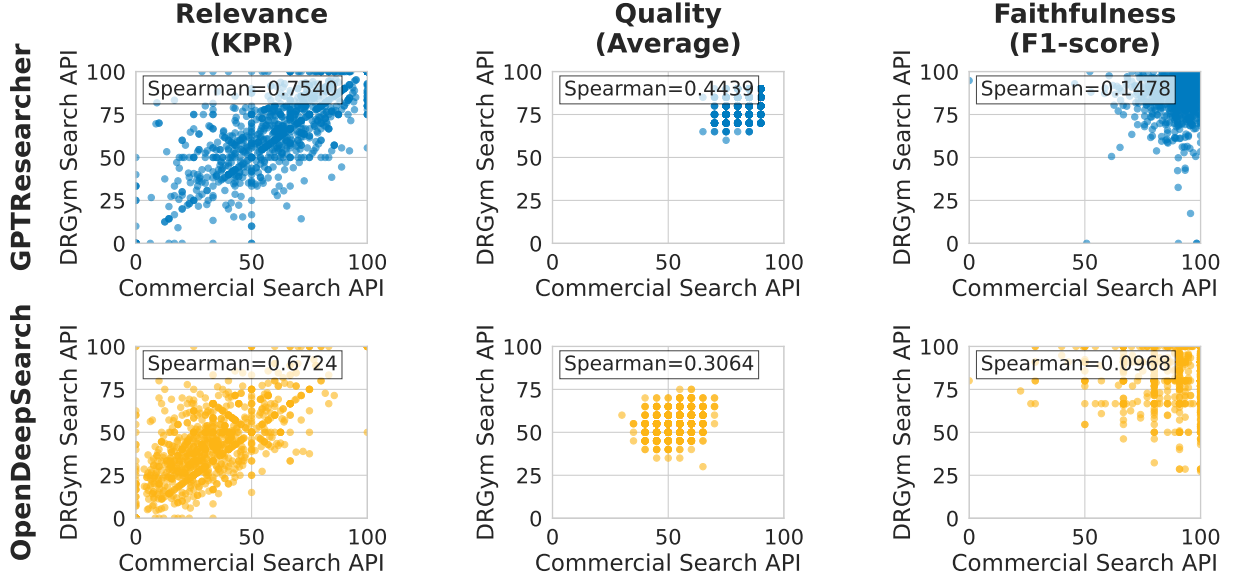


Figure 3: Query-level correlation across metrics, when changing between commercial search (i.e., Serper) and our API.

Table 5: LLM versus human agreement (Cohen’s κ).

Metric	κ
KPR	0.72
Citation Precision	0.86
Citation Recall	0.86
Clarity	0.89
Insightfulness	0.84

showing the percentage of times each system was preferred over another based on human annotations. The KPC metric was excluded from the analysis in Table 5 due to a high rate of LLM-judged ties, leaving insufficient samples for analysis. The results demonstrate substantial agreement and the same trend regarding relative model performance. This validates our LLM-as-a-judge evaluation protocol, strengthening confidence in the benchmark results presented in the previous sections.

6.7 Qualitative Analysis

To understand where key-point recall breaks down, we performed a qualitative analysis on GPTResearcher, the best-performing open-source system in our benchmark. We took the 100 queries with the lowest key-point recall and inspected them manually. In most cases, the system produced an accurate report but missed the facet or framing captured in the gold key-points. This typically occurred when the query was underspecified relative to the more specific information need reflected in the source documents. We group them into three categories.

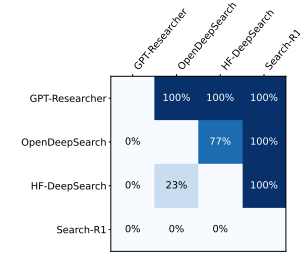


Figure 4: Win rates comparing pairs of systems as judged by human preferences.

Multi-facet coverage gaps account for 78.5% of missing key-points and occur when a broad query admits several plausible readings. GPTResearcher commits to the most salient interpretation and does not hedge across the remaining ones. On the query “*does the american population wish more autonomy in their work?*”, the system produced a thorough report on workplace autonomy covering flexibility, remote work, and work-life balance. It did not surface gold key-points tied to autonomy in adjacent senses, such as undocumented workers and immigration policy, physician ownership of private practice, patient autonomy in medical ethics, and generational framing.

Lens mismatches account for 17.9% of missing key-points and arise when the report answers at a reasonable but different level of abstraction than the gold sources. When asked “*what makes an athlete elite?*”, the system discussed psychological traits, training environments, coaching relationships, and genetic factors, all defensible aspects of elite performance. The gold key-points were drawn from a source centered on neuromuscular mechanisms such as brain-muscle communication, motor control centers, and training-induced neural adaptation. The report is thematically correct but never engages this framing, which the query does not single out as the intended lens.

Domain misinterpretations account for 3.6% of missing key-points and are the closest to a genuine failure, since ambiguous terminology leads the system to research a different topic. For the query “*what role did indians play in the wars for empire?*”, GPTResearcher interpreted “Indians” as people from India and reported on Indian troops in British imperial wars, whereas the gold document focuses on “Indians” as Native American and Indigenous peoples and discusses Aztec, Inca, and Maya societies and Indigenous alliances with European powers in eighteenth-century North American conflicts.

Taken together, over 90% of the unrecovered key-points fall into the first two categories and stem from query underspecification.

Resolving them would require the system to determine which interpretation, or whose information need, a query targets, which a single zero-shot pass cannot do. Multi-turn clarification and personalization address this directly, as in recent work that adds mid-process steering and persona modeling to long-horizon research [1]. Our evaluation, like GPTResearcher itself, models a single-turn setting and does not capture such interaction.

7 Conclusions and Future Work

DeepResearchGym. We offer a reproducible sandbox for developing and benchmarking deep research systems, providing a stable cost-effective alternative to commercial search APIs. We aim to depart from full reliance on black-box commercial solutions, where algorithms and data are not fully known. In contrast, DEEPRSEARCHGYM is open in terms of data and algorithms, allowing the community to reproduce and measure scientific progress in deep research agents. We release all components, including the retrieval API, the query generation framework, the evaluation benchmark, and the training and evaluation code, at <https://www.deeprsearchgym.ai>.

Use cases. We empirically demonstrate that DEEPRSEARCHGYM serves as a reliable research-grade complement to commercial retrieval infrastructures. Training-wise, we show that agents trained exclusively within DEEPRSEARCHGYM generalize effectively to commercial search at inference time, validating the framework as a practical training environment. Beyond training, a benchmarking experiment on long-report generation shows that systems maintain comparable performance when transitioning from proprietary APIs to our transparent environment, confirming retrieval fidelity. By isolating system behavior from fluctuating retrieval conditions and API expenses, DEEPRSEARCHGYM provides a foundation for reproducible and accessible research in deep research systems.

Follow-up Work. We are committed to expanding DEEPRSEARCHGYM's coverage to recent web corpora, such as newer FineWeb crawls, enabling evaluation of time-sensitive queries and emerging topics. Moreover, the integration of domain-specific benchmarks may further support assessment in high-stakes contexts such as healthcare or law, where retrieval precision and factual reliability are critical. For reproducibility, such extensions will be provided in novel versioned endpoints.

Acknowledgments

The Portuguese researchers were supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (i.e., the Center For Responsible AI), and also by the Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (<https://doi.org/10.54499/UID/50021/2025>), and UID/PRR/50021/2025 (<https://doi.org/10.54499/UID/PRR/50021/2025>). João Coelho was additionally supported through the Ph.D. scholarship with reference PRT/BD/153683/2021, under the CMU Portugal Program.

References

- [1] Lin Ai, Victor S. Bursztyn, Xiang Chen, Julia Hirschberg, and Saayan Mitra. 2026. An Interactive Paradigm for Deep Research. ArXiv abs/2605.24266.
- [2] Salaheddin Alzubi, Creston Brooks, Purva Chiniya, Edoardo Contente, Chiara von Gerlach, Lucas Irwin, Yihan Jiang, Arda Kaz, Windsor Nguyen, Sewoong Oh, Himanshu Tyagi, and Pramod Viswanath. 2025. Open Deep Search: Democratizing Search with Open-source Reasoning Agents. ArXiv abs/2503.20201.
- [3] BAAI. 2024. bge-large-en-v1.5. <https://huggingface.co/BAAI/bge-large-en-v1.5> Hugging Face Model Page.
- [4] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. In *Conference on Language Modeling, COLM*.
- [5] Michael Bendersky and W. Bruce Croft. 2009. Analysis of Long Queries in a Large Scale Search Log. In *Workshop on Web Search Click Data, WSDM*.
- [6] Google Blog. 2024. Generative AI in Search: Let Google Do the Searching for You. <https://blog.google/products/search/generative-ai-google-search-may-2024/> Accessed: 2025-05-12.
- [7] Google Blog. 2025. Why we're taking legal action against SerpApi's unlawful scraping. <https://blog.google/innovation-and-ai/technology/safety-security/serpapi-lawsuit/> Accessed: 2026-01-28.
- [8] Francesco Bonchi, Carlos Castillo, Debora Donato, and Aristides Gionis. 2008. Topical query decomposition. In *International Conference on Knowledge Discovery and Data Mining, KDD*.
- [9] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (2002).
- [10] Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifmoghadam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. 2025. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. ArXiv abs/2508.06600.
- [11] Assaf Elovic. 2025. GPT Researcher: LLM-based Autonomous Agent for Deep Research. <https://github.com/assafelovic/gpt-researcher>
- [12] Jiaxuan Gao, Wei Fu, Mingyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. 2025. Beyond Ten Turns: Unlocking Long-Horizon Agentic Search with Large-Scale Asynchronous RL. ArXiv abs/2508.07976.
- [13] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A Survey on LLM-as-a-Judge. ArXiv abs/2411.15594.
- [14] Ruijun Han, Yanfei Chen, Zoey CuiZhu, Lesly Miculicich, Guan Sun, Yuanjun Bi, Weiming Wen, Hui Wan, Chunfeng Wen, Solène Maitre, George Lee, Vishy Tirumalashetty, Emily Xue, Zizhao Zhang, Salem Haykal, Burak Gokturk, Tomas Pfister, and Chen-Yu Lee. 2025. Deep Researcher with Test-Time Diffusion. ArXiv abs/2507.16075.
- [15] Shengding Hu et al. 2024. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. In *Conference on Language Modeling, COLM*.
- [16] HuggingFace. 2025. Open Deep Research. https://github.com/huggingface/smolagents/tree/main/examples/open_deep_research
- [17] Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. ArXiv abs/2503.09516.
- [18] Jiahe Jin, Abhijay Paladugu, and Chenyan Xiong. 2026. Beneficial Reasoning Behaviors in Agentic Search and Effective Post-training to Obtain Them. In *Annual Meeting of the Association for Computational Linguistics, ACL*.
- [19] JinaAI. 2024. jina-embeddings-v3. <https://huggingface.co/jinaai/jina-embeddings-v3> Hugging Face Model Page.
- [20] Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananeey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqi. 2025. Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation. In *Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*.
- [21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Annual Conference on Neural Information Processing Systems, NeurIPS*.
- [22] Weizhen Li et al. 2025. Chain-of-Agents: End-to-End Agent Foundation Models via Multi-Agent Distillation and Agentic RL. ArXiv abs/2508.13167.
- [23] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic Search-Enhanced Large Reasoning Models. ArXiv abs/2501.05366.
- [24] Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capability. ArXiv abs/2504.21776.
- [25] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Conference on Empirical Methods in Natural Language Processing, EMNLP*.
- [26] Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. GAEA: A Benchmark for General AI Assistants. In *International Conference on Learning Representations, ICLR*.
- [27] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Conference on Empirical Methods in Natural Language Processing, EMNLP*.
- [28] Niklas Muennighoff. 2022. SGPT: GPT Sentence Embeddings for Semantic Search. ArXiv abs/2202.08904.

- [29] Xuan-Phi Nguyen, Shrey Pandit, Revanth Gangi Reddy, Austin Xu, Silvio Savarese, Caiming Xiong, and Shafiq Joty. 2025. SFR-DeepResearch: Towards Effective Reinforcement Learning for Autonomously Reasoning Single Agent. *ArXiv abs/2509.06283*.
- [30] Jingjie Ning, João Coelho, Yibo Kong, Yunfan Long, Bruno Martins, João Magalhães, Jamie Callan, and Chenyan Xiong. 2026. Agentic Search in the Wild: Intents and Trajectory Dynamics from 14M+ Real Search Requests. In *International Conference on Research and Development in Information Retrieval, SIGIR*.
- [31] OpenAI. 2025. Introducing Deep Research. <https://openai.com/index/introducing-deep-research/> OpenAI Blog. May 11, 2025.
- [32] OpenBMB. 2024. MiniCPM-Embedding-Light. <https://huggingface.co/openbmb/MiniCPM-Embedding-Light> Hugging Face Model Page.
- [33] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Rich Information. In *International Conference on Research and Development in Information Retrieval, SIGIR*.
- [34] Guilherme Penedo, Hynek Kydlicek, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin A. Raffel, Leandro von Werra, and Thomas Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *Annual Conference on Neural Information Processing Systems NeurIPS*.
- [35] Perplexity AI. 2025. Introducing Perplexity Deep Research. <https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research> Perplexity Blog. Accessed May 11, 2025.
- [36] Long Phan et al. 2025. Humanity's Last Exam. *ArXiv abs/2501.14249*.
- [37] Zehan Qi, Rongwu Xu, Zhijiang Guo, Cunxiang Wang, Hao Zhang, and Wei Xu. 2024. LONG²RAG: Evaluating Long-Context & Long-Form Retrieval-Augmented Generation with Key Point Recall. In *Findings of the Association for Computational Linguistics, EMNLP*.
- [38] Corby Rosset, Ho-Lam Chung, Guanghui Qin, Ethan C. Chau, Zhuo Feng, Ahmed Awadallah, Jennifer Neville, and Nikhil Rao. 2024. Researchy Questions: A Dataset of Multi-Perspective, Decompositional Questions for LLM Web Agents. *ArXiv abs/2402.17896*.
- [39] Rishiraj Saha Roy, Smith Agarwal, Niloy Ganguly, and Monojit Choudhury. 2016. Syntactic Complexity of Web Search Queries through the Lenses of Language Models, Networks and Users. *Information Processing and Management* 52, 5 (2016).
- [40] Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation. In *Annual Conference on Neural Information Processing Systems, NeurIPS*.
- [41] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*.
- [42] Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-Solve-Merge Improves Large Language Model Evaluation and Generation. In *Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*.
- [43] Serper. 2025. Serper.dev Search API. <https://serper.dev> Accessed May 11, 2025.
- [44] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *ArXiv abs/2402.03300*.
- [45] Dingfeng Shi, Jingyi Cao, Qianben Chen, Weichen Sun, Weizhen Li, Hongxuan Lu, Fangchen Dong, Tianrui Qin, King Zhu, Minghao Liu, Jian Yang, Ge Zhang, Jiaheng Liu, Changwang Zhang, Jun Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2025. TaskCraft: Automated Generation of Agentic Tasks. *ArXiv abs/2506.10055*.
- [46] Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement Learning. *ArXiv abs/2503.05592*.
- [47] Suhas Jayaram Subramanya, Devvrit, Harsha Vardhan Simhadri, Ravishankar Krishnaswamy, and Rohan Kadekodi. 2019. DiskANN: Fast Accurate Billion-point Nearest Neighbor Search on a Single Node. In *Annual Conference on Neural Information Processing Systems, NeurIPS*.
- [48] Tavily. 2025. Tavily Search API. <https://tavily.com> Accessed May 11, 2025.
- [49] Bing Team. 2025. Introducing Copilot Search in Bing. <https://blogs.bing.com/search/April-2025/Introducing-Copilot-Search-in-Bing> Accessed: 2026-01-28.
- [50] Nandan Thakur, Ronak Pradeep, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Support Evaluation for the TREC 2024 RAG Track: Comparing Human versus LLM Judges. *ArXiv abs/2504.15205*.
- [51] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Conference on Neural Information Processing System, NeurIPS*.
- [52] Lei Wang, Wanyu Xu, Yihui Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In *Annual Meeting of the Association for Computational Linguistics, ACL*.
- [53] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. In *Annual Conference on Neural Information Processing Systems, NeurIPS*.
- [54] Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025. WebWalker: Benchmarking LLMs in Web Traversal. In *Annual Meeting of the Association for Computational Linguistics, ACL*.
- [55] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A Critical Evaluation of Evaluations for Long-form Question Answering. In *Annual Meeting of the Association for Computational Linguistics, ACL*.
- [56] An Yang et al. 2025. Qwen3 Technical Report. *ArXiv abs/2505.09388*.
- [57] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. DeepResearcher: Scaling Deep Research via Reinforcement Learning in Real-world Environments. *ArXiv abs/2504.03160*.
- [58] Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhao Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S. Yu. 2024. Trustworthiness in Retrieval-Augmented Generation Systems: A Survey. *ArXiv abs/2409.10102*.