DeepSearchQA: Bridging the Comprehensiveness Gap for Deep Research Agents

Nikita Gupta[*,◇], Riju Chatterjee[*,♠], Lukas Haas[*,◇], Connie Tao[*,◇], Andrew Wang[♡], Chang Liu[♣], Hidekazu Oiwa[◇], Elena Gribovskaya[◇], Jan Ackermann[◇], John Blitzer[◇], Sasha Goldshtein[♣] and Dipanjan Das[◇]
[*]Equal Contribution, [◇]Google DeepMind, [♠]Google Search, [♡]Kaggle, [♣]Google Research

We introduce *DeepSearchQA*, a 900-prompt benchmark for evaluating agents on difficult multi-step information-seeking tasks across 17 different fields. Unlike traditional benchmarks that target single-answer retrieval or broad-spectrum factuality, DeepSearchQA features a dataset of challenging, hand-crafted tasks designed to evaluate an agent's ability to execute complex search plans to generate exhaustive answer lists. This shift in design explicitly tests three critical, yet under-evaluated capabilities: 1) systematic collation of fragmented information from disparate sources, 2) de-duplication and entity resolution to ensure precision, and 3) the ability to reason about stopping criteria within an open-ended search space. Each task is structured as a causal chain, where discovering information for one step is dependent on the successful completion of the previous one, stressing long-horizon planning and context retention. All tasks are grounded in the open web with objectively verifiable answer sets. Our comprehensive evaluation of state-of-the-art agent architectures reveals significant performance limitations: even the most advanced models struggle to balance high recall with precision. We observe distinct failure modes ranging from premature stopping (under-retrieval) to hedging behaviors, where agents cast an overly wide net of low-confidence answers to artificially boost recall. These findings highlight critical headroom in current agent designs and position DeepSearchQA as an essential diagnostic tool for driving future research toward more robust, deep-research capabilities. DeepSearchQA can be found at https://www.kaggle.com/benchmarks/google/dsqa/leaderboard.

1. Introduction

The field of artificial intelligence is witnessing a paradigm shift, marked by the rapid transition from static Large Language Models (LLMs) to autonomous web agents designed to interact with dynamic and complex environments to achieve specific goals. This "agentic revolution" demands sophisticated capabilities like planning, memory management and tool use. As these systems are increasingly integrated into production-level workflows, tasked with everything from simple data retrieval to complex, multi-step research, the necessity for robust, realistic evaluation methodologies has become paramount. Effective benchmarks are crucial not only for measuring progress but for identifying critical failure modes in these open-ended domains. However, the development of agent capabilities is rapidly outpacing current evaluation methodologies, leading to a significant evaluation bottleneck where existing benchmarks are either saturated or too contrived to represent real-world user needs.

1.1. The Prevailing Paradigm: Single-Answer Verification

The previous evaluation methods such as TruthfulQA (Lin et al., 2022), HaluEval (Li et al., 2023), and FELM (Chen et al., 2023), or precision benchmarks like SimpleQA (Haas et al., 2025; Wei et al., 2024a) have successfully established rigorous standards for factuality and single-answer retrieval tasks such as *"What is the capital of France?"*. This single-answer format was a deliberate and highly effective strategy. It enabled low-cost, scalable, and objective automated grading, removing the subjectivity and expense associated with evaluating long-form, generative answers—a massive challenge highlighted by recent

---

Table 1 | Representative examples of user information needs in *DeepSearchQA*. Each row maps a specific domain profile to a complex retrieval task, illustrating the diversity of topics and aggregation requirements.

Profile	Prompt
The Epidemiologist	Between 2019 and 2020, which states had a death rate (the number of deaths per 100,000 total population) between 229.8 - < 268.1 that was inflicted by the leading cause of death in the US during 2019 and 2020 per the CDC?
The Video Gamer	I love video games and I'm doing a bit of an analysis into some of the highest selling hardware. Out of the 'PlayStation', 'PlayStation 2', 'PlayStation 3', 'PSP', 'Nintendo DS', 'Nintendo Wii', 'Xbox 360' and 'Game Boy Advance', I want to know which ones sold over 100 million hardware units, include all variations of each that fall under the same family of releases (Use company websites and press releases for this data). Of those which sold more than 100 million units and aren't portable handheld devices, which of these had a 32 bit CPU? And finally, for those devices which satisfy all my constraints, tell me which of their parent company's had assets totalling over 1.76 trillion Yen by the end of the financial year 2010 (March 31st 2010), (use official company financial reports for this data).
The Safety Auditor	According to the NHTSA, which US states had a fatality rate per 100 million vehicle miles traveled below 1 in two or more consecutive years between 2000 and 2005.
The Demographer	Among Canadian provinces and territories that had more than 30 small population centers in 2016, which experienced positive GDP growth and had an Aboriginal identity population exceeding 50,000 in the same year? Use Statistics Canada as the source for all data.

work on hallucination and attribution (Rashkin et al., 2022; Zhao et al., 2024). While researchers have developed metrics for long-form consistency—such as FActScore (Bishop et al., 2023; Min et al., 2023), LongFact (Wei et al., 2024b), VERISCORE (Song et al., 2024), and FactAlign (Huang and Chen, 2024)—these often require expensive human-in-the-loop verification or complex LLM-based judges (Liu et al., 2024; Ramprasad and Wallace, 2024). Similarly, claim verification frameworks like TRUE (Honovich et al., 2022), MiniCheck (Tang et al., 2024), and CoverBench (Jacovi et al., 2024) focus on grounding specific claims rather than evaluating the completeness of a research trajectory.

With the release of reasoning-intensive models (Anthropic, 2025; DeepSeek-AI et al., 2025; Gemini Team, Google, 2025; OpenAI, 2024, 2025c), agentic evaluation requires assessing the ability to reason over external information, formulate long-horizon plans, and synthesize findings from unstructured web sources. Previous attempts to address this, such as Humanity's Last Exam (HLE) (Phan et al., 2025), GAIA (Mialon et al., 2023), and OpenAI's BrowseComp (Wei et al., 2025), have pushed the difficulty cieling by introducing expert-level questions that resist simple retrieval. However, even these advanced benchmarks largely retain the single-answer format to ensure objective scoring. Even benchmarks designed for dynamic or search-augmented settings, such as FreshLLMs (Vu et al., 2023) and RealTime QA (Kasai et al., 2024), often rely on atomic answers that do not capture the complexity of broad information gathering.

This single-answer paradigm was a valuable scaffold, isolating core competencies like navigation and search strategy (Miroyan et al., 2025). However, it incentivizes precision-first distinct search trajectories —finding the needle in the haystack—rather than the recall-oriented exhaustive research required in many real-life user information-seeking journeys (Du et al., 2025; OpenAI, 2025a).

1.2. The Comprehensiveness Gap: Beyond Finding an Answer

The limitation of the single-answer approach reveals what we term the *Comprehensiveness Gap* in current agent evaluation. Many real-world information-seeking tasks are not satisfied by retrieving a single data point, as seen in traditional datasets like Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), or HotpotQA (Yang et al., 2018). Instead, they require more complex research tasks such as *"List all companies in the semiconductor sector with a P/E ratio under 20 and a presence in Southeast Asia."* or *"Identify all clinical trials for mRNA vaccines initiated in 2024."*
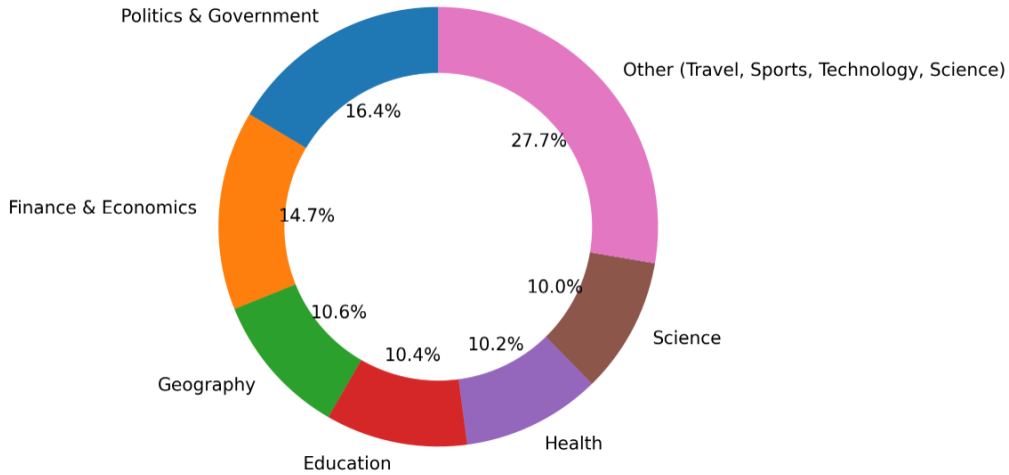
Figure 1 | Distributions of prompt categories in the benchmark. This figure shows the relative sizes of the five main categories and another category covering all other types of prompts. The class sizes are well balanced, creating a diverse benchmark.

This shift from simple retrieval to agentic information retrieval (Zhang et al., 2024) necessitates a fundamentally different cognitive profile. It requires three higher-order capabilities that existing benchmarks fail to isolate:

1. First, the agents must demonstrate *Systematic Collation*, the ability to visit disparate sources — often hundreds — and collate a master list where no single source contains the comprehensive information. Recent work on unified retrieval evaluation, such as Fact-Fetch-Reason (Krishna et al., 2025) begins to address the agentic nature of information retrieval, but lacks the exhaustive scope required here.

2. Second, agents must perform *Entity Resolution (De-duplication)*. This is the ability to identify when two retrieved entities are identical despite varying surface forms — a complex task in structured reasoning (Li et al., 2024) often overlooked in implicit reasoning benchmarks (Geva et al., 2021). In deep research context, failure to resolve entities leads to inflated lists and degraded precision, a common failure mode we observe in our experiments.

3. Third, and perhaps most critically, agents must reason about *Stopping Criteria* under epistemic uncertainty. The agent must determine when a search is complete without an explicit termination signal (Yang et al., 2024). This challenges the agent to distinguish between absence of evidence ("I have not found it yet.") and evidence of absence ("It does not exist.").

1.3. DeepSearchQA Core Contributions

The primary contribution of *DeepSearchQA* is to shift the evaluation paradigm from precision-based retrieval to exhaustive answer set generation. The benchmark requires agents to conduct deep, autonomous browsing operations to generate a complete, verifiable set of all possible answers for a given query, rather than a single data point. In line with the principles of simplicity and verifiability espoused by SimpleQA (Haas et al., 2025; Wei et al., 2024a), Facts (Cheng et al., 2025), and BrowseComp (Wei et al., 2025), DeepSearchQA employs a strictly outcome-based evaluation methodology. Performance is judged solely on the completeness (recall) and correctness (precision) of the final set submitted by the agent, rather than the search trajectory used to obtain it. This approach encourages architectural diversity while serving as a rigorous functional test. To achieve a high F1 score, an agent is effectively forced to master the trade-off between exploration (casting a wide net) and exploitation (verifying candidates), navigating the open web without a predefined map.

We will maintain a live *DeepSearchQA* leaderboard tracking performance of different models. The leaderboard will remain open to new model submissions. Evaluation will be done by Kaggle.

Task Complexity	Subtypes	Easy Example	Hard Example
Structured Retrieval (The "Search")	Multi-step search strategy, Obscure info	**Linear Fact Retrieval** "What is the capital of France?" *(Single step, high redundancy)*	**Dependency Graph Retrieval** "Which cities in the state where LaMotte established a fort have Amtrak stations?"
Context Management (The "Assembly")	Full page / doc comprehension	**Short-Context Extraction** "Summarize the first paragraph of this news article."	**Massive-Context "Needle"** "Based on these 50 uploaded PDF case files, list every instance where a judge cited 'Section 404' regarding a liability cap."
Logical Reasoning (The "Thinking")	Analytical reasoning, Information synthesis	**Simple Aggregation** "List the 5 biggest companies by market cap."	**Inference & Constraint Solving** "Identify the US state that has the highest number of Fortune 500 headquarters relative to its population, excluding states that have zero corporate income tax."

Table 2 | Taxonomy for benchmarking task complexity across domains. The second column describes which subtypes of common cognitive reasoning problems are encompassed by each task. The last two columns compare easy against hard versions of the same examples.

2. DeepSearchQA: Dataset and Taxonomy

The *DeepSearchQA* dataset comprises a diverse set of questions curated by expert data annotators from various sources to reflect realistic, high value user queries. These were rigorously filtered to focus on objective, information-seeking tasks where the ground truth is definitive.

Dataset Statistics and Domains. The benchmark contains 900 prompts paired with ground-truth answer sets. The questions span a broad distribution of domains—including Politics and Government, Finance and Economics, Science, Health, History, Geography, and Media—ensuring that agents are tested on their ability to generalize across varying web structures and content types. Figure 1 presents the distribution of the domains in the dataset and Table 1 shows different examples from our curated dataset. A critical design choice was to ensure that all prompts are time-anchored or reference static data sources (e.g., "*According to the 2020 Census...*"). This minimizes the drift inherent in live web benchmarks, where the ground truth changes over time.

Answer Types. The ground-truth answer space is categorized into two distinct types:

- **Single Answer:** The answer is a unique entity or value (e.g., a specific date or name). While similar to traditional QA, these require deep research due to the obscurity of the requested data and possible conflicting evidences on the web.
- **Set Answer:** The answer is a collection of items. This includes *Enumeration* (listing all items matching a constraint) and *Composite* responses (answering multiple sub-questions).

Quality Verification Protocol. To ensure ground truth accuracy, we implemented a standardized three-phase verification protocol. The process included a 1) *Independent Research Phase*, where three

Full Prompt	Task Complexity
Relocation Planner I'm looking to move house to a new city. Which cities have an average house price of less than £200k according to the 2023 Hometrack 65 city price index? Of those cities, filter down the list to the top 5 English cities with the highest percentage of green space according to Ordnance Survey's 2023 Open Greenspace Data. Filter these cities down to the top 3 by percentage of people economically active and in employment according to the Office for National Statistics annual population survey for the Jan 2022-Dec 2022 period. I also drive and I find the new initiatives make planning car journeys convoluted. Therefore, filter the list down further to cities without a clean air zone introduced before January 2024.	Requires multi-step structured retrieval strategies while maintaining context state across four distinct domain searches to narrow a solution set. Filter Housing (<£200k) → Rank Green Space (Top 5) → Rank Employment (Top 3) → Exclude Clean Air Zones
Vaccination Study I am researching small countries to use for a case study on vaccinations. Can you give me a complete list of all countries with a population between 1 and 2 million people, a life expectancy of 75 years or more, and a measles immunization rate of at least 85%? Use World Bank data from 2023, do not include any country that doesn't have immunization data from that year, and just give me a comma separated list of countries. No additional text please.	Requires traversing a dataset to find entities satisfying a complex dependency graph of attributes (Pop ∩ LifeExp ∩ Vax). Filter Population (1-2M) → Filter Life Expectancy (>75) → Filter Immunization (>85%) → Extract Intersection
Financial Trend Hunter Refer to macrotrends.net for NVIDIA's historical annual stock price data and to worldometers.info for United States GDP percentage change. For the years 2020, 2021, 2022, and 2023, tell me which of these saw NVIDIA's stock price grow by more than 125% (annual % change) and the GDP of the United States grow by more than 2.5%.	Requires synthesizing information from disparate sources and solving mathematical constraints ('Stock > 125%' AND 'GDP > 2.5%'). Retrieve Macrotrends Data → Retrieve Worldometer Data → Calculate Annual Growth → Intersect Thresholds

Table 3 | Representative DeepSearchQA tasks. The analysis illustrates the complexity of queries requiring structured retrieval with constraint verification, context management and cross-domain synthesis.

reviewers attempted to independently research the solution without access to the ground truth. This was followed by 2) *Verification and Comparison*, where the independent answer was cross-referenced with the original ground-truth answer given by the prompt curator. Any discrepancies triggered a final 3) *Conflict Resolution Phase*, to determine if the original ground-truth answer required updating, the reviewer was incorrect, or the prompt itself was ambiguously flawed. All ambiguous prompts were filtered at this stage.

Task Complexity Taxonomy. We categorize tasks based on the cognitive and operative demands they place on the agent. This taxonomy helps diagnose specific weaknesses in agent architectures.

- **Structured Retrieval ("The Search"):** This requires devising multi-step search strategies to retrieve obscure, niche information to fulfill the user need.
- **Context Management ("The Assembly"):** This involves ingesting and synthesizing large volumes of information of potentially different formats, where the key challenge is about effectively managing context window limitations.
- **Logical Reasoning ("The Thinker"):** This requires abstract deduction, planning, conflict resolution, and infer latent information from incomplete data.

Table 2 shows one example for each of the tasks and illustrates the difference between easy examples from previous benchmarks compared to hard examples which we include in our benchmark and Table 3 illustrates the required reasoning to solve prompts from our new benchmark.

3. Evaluation Methodology

The evaluation of DeepSearchQA requires a methodology that can robustly and fairly assess the quality of a set-based answer format, where the order of items is irrelevant and semantic variations in answers are common.

3.1. Formal Evaluation Metrics

To formally evaluate performance, we adopt standard metrics from the field of information retrieval as well as introduce several categorical metrics. In the following definitions, let S_i be the set of distinct answers submitted by the agent for a given prompt i, and G_i be the ground-truth set of correct answers for that prompt. We utilize a two-tiered evaluation approach: *Continuous Metrics* for granular performance analysis, and *Categorical Classification* for strict success rates.

Continuous Metrics. We report standard retrieval metrics to quantify the trade-off between accuracy and exhaustiveness. These metrics are calculated for each individual prompt and then averaged over the entire evaluation set. *F1-Score* (F_1) serves as the primary ranking metric, providing a balanced measure of performance by calculating the harmonic mean of Precision (P) and Recall (R).

For a specific prompt i:

$$F_{1,i} = 2 \cdot \frac{P_i \cdot R_i}{P_i + R_i} \tag{1}$$

Here, precision ($P_i = \frac{|S_i \cap G_i|}{|S_i|}$) represents the accuracy of the submitted items, while Recall ($R_i = \frac{|S_i \cap G_i|}{|G_i|}$) measures the exhaustiveness of the submission against the ground truth.

- For single-answer tasks: The F1 score is equivalent to Exact Match accuracy (binary 1.0 or 0.0).
- For set-answer tasks: The F1 score ensures performance reflects both completeness and accuracy. This metric strictly penalizes agents that attempt to maximize Recall by simply outputting a large number of guesses (hallucinations or "drift").

Categorical Classification. We also report strict, binary categorizations of each response into one of four disjoint categories based on the set relationship between S_i and G_i:

1. **Fully Correct ($S_i = G_i$):** A response is fully correct if and only if the submitted set is semantically identical to the ground-truth set. The agent must identify all correct answers while including zero incorrect answers.

   - Single-answer tasks: The agent provides the exact, unique golden entity required.
   - Set-answer tasks: The agent retrieves the complete, exhaustive list of items ($R_i = 1.0$) without hallucinating any extra items ($P_i = 1.0$).

2. **Fully Incorrect ($S_i \cap G_i = \emptyset$):** A response is fully incorrect if the submitted set contains zero correct items. The intersection between the submitted answer and the ground truth is empty.

   - Single-answer tasks: The agent provides a wrong answer or claims no answer exists.
   - Set-answer tasks: The agent fails to retrieve even a single valid item from the ground truth answer set. This suggests a complete failure of the search strategy or a fundamental misunderstanding of the prompt.

3. **Partially Correct ($\emptyset \neq (S_i \cap G_i) \subset G_i$):** A response is partially correct, if it contains some but not all correct answers and potentially some not contained by the ground truth. This error mode only exists for set-answer tasks.

4. **Correct with Extraneous Answers (**$G_i \subset S_i$**):** A response falls into this category if the agent successfully identifies all items from the ground-truth answer set ($R_i = 1.0$) but also includes one or more incorrect items ($P_i < 1.0$). This represents the *hedging* failure mode.

   - Single-answer tasks: This is treated as a failure. Providing multiple candidates when a unique entity is required implies a failure to disambiguate.
     - Example (Conflict Avoidance): *Q: "Who won the 1994 World Cup?" (GT: Brazil)* → Agent: "Brazil and Italy." The agent lists both the winner and runner-up rather than committing to one.
     - Example (Popularity Bias): *Q: "What is the capital of Australia?" (GT: Canberra)* → Agent: "Canberra or Sydney." The agent includes the most famous city alongside the actual capital to mitigate the risk of user error.
   - Set-answer tasks: The agent identifies all correct items ($R_i = 1.0$) but fails the *stopping criteria*, effectively guessing extra items.
     - Example (Classification Drift): *Q: "List the planets in the Solar System."* → Agent includes Pluto alongside the correct eight.
     - Example (Association): *Q: "Which US states border Maine?"* → Agent: "New Hampshire and Vermont." The agent hallucinates items that are semantically or geographically close to the target.

3.2. Automated Evaluation Pipeline

We employ an automated *LLM-as-a-Judge* pipeline to determine the semantic equivalence of extracted answers. For every item $s \in S_i$, the judge determines if s is semantically equivalent to any item $g \in G_i$. For evaluation purposes, we use Gemini 2.5 Flash (Gemini Team, Google, 2025) in a zero-shot setting using the prompt provided in Appendix A.

4. Results and Analysis

Table 4 presents the comprehensive evaluation performance on a suite of state-of-the-art *Deep Research* agents and reasoning models. The results highlight the scaling behavior of agentic capability and the persistent difficulty of the *Comprehensiveness Gap*. We report four primary indicators: the percentage of *Fully Correct* responses ($S = G$, where the model retrieved the exact ground-truth set), *Fully Incorrect* responses ($S \cap G = \emptyset$, where no relevant items were found), and the aggregate *F1-Score*, which balances precision and recall. We also report *Correct with Extraneous Answers* where models retrieve the correct answer but fail to filter out non-relevant items.

State-of-the-Art in Deep Research. *Gemini Deep Research Agent* and *GPT-5 Pro High Reasoning* (OpenAI, 2025b) establish the state-of-the-art on this benchmark, although *Gemini Deep Research Agent* likely does so more efficiently, judging by the significantly lower cost[1]. The results reveal a clear hierarchy: Deep Research Agents outperform their standalone reasoning model counterparts. This confirms that while reasoning models provide the raw intelligence, the iterative loop of an agent is required to close the *Comprehensiveness Gap*.

Notably, *GPT-5 Pro High Reasoning* and *Gemini Deep Research Agent* exhibit different performance profiles. While both have statistically comparable fully correct rates (66.09% vs 65.18%), the *Gemini Deep Research Agent* minimizes catastrophic failures more effectively, achieving the lowest *Fully Incorrect* rate of 9.95% vs 14.13% for GPT-5 Pro high reasoning. Table 5 illustrates common failure modes exhibited by the top-performing models. Notably, these failures manifest at distinct stages of the information processing pipeline. In the first instance, the *Gemini Deep Research Agent* successfully

---

[1]Gemini Deep Research Agent pricing vs. GPT-5 Pro pricing.

Table 4 | Main results on DeepSearchQA, independently evaluated by Kaggle. We report the relative number of Fully Correct, Fully Incorrect, and Correct with Extraneous Answers. Additionally, we present the F1-Score achieved by each method. The small number to the right of each metric is the 95 percent confidence interval.

#	Model	Fully correct ↑	Fully incorrect ↓	Correct with Extraneous Answers ↓	F_1 ↑
1	Gemini Deep Research Agent	**66.09** ±3.16	**9.95** ±2.00	10.30 ±2.03	**81.90**
2	GPT-5 Pro High Reasoning	65.18 ±3.11	14.13 ±2.28	8.12 ±1.79	78.98
3	GPT-5 High Reasoning	59.41 ±3.21	19.91 ±2.61	6.56 ±1.62	73.24
4	Gemini 3 Pro Preview	56.56 ±3.24	12.78 ±2.18	9.89 ±1.95	76.86
5	o3 Deep Research	44.24 ±3.27	20.09 ±2.64	11.74 ±2.12	66.45
6	o4 Mini Deep Research	40.36 ±3.21	24.19 ±2.80	7.80 ±1.76	61.76
7	Gemini 2.5 Flash	25.92 ±2.86	45.27 ±3.25	5.90 ±1.54	42.99
8	Claude 4.5 Opus	24.01 ±2.89	50.66 ±3.39	4.18 ±1.36	40.20
9	Claude 4.5 Sonnet	16.04 ±2.40	64.25 ±3.13	2.90 ±1.10	27.85
10	Claude 4.5 Haiku	12.78 ±2.18	71.00 ±2.96	1.89 ±0.89	22.24

aggregates the raw data but fails to synthesize the final estimate. In the second, *GPT-5 Pro High Reasoning* encounters an unreadable document and fails to recover, effectively halting the process. Finally, the third example demonstrates a constraint violation, where the model fails to filter data points according to the prompt's requirements, indicating a breakdown in the agent's stopping criteria.

The mid-tier consists of *o3 Deep Research* (Fully Correct: 44.24%) and *o4 Mini Deep Research* (Fully Correct: 40.36%). Their Fully Incorrect rates rise to 20.09% and 24.19% respectively, showing rapid degradation in the ability to maintain coherent long-horizon search strategies.

Scaling Behavior and the Reasoning Threshold. The *F1-Score* and *Fully Incorrect* metrics highlight a hard *reasoning threshold* required for autonomous research. While the gap between the top models is relatively narrow, there is a precipitate drop-off for smaller reasoning models. *Gemini 2.5 Flash*, for instance, drops to an F_1 of 42.99%, roughly half that of the leader. More critically, the *Fully Incorrect* rate spikes to 45.27%—nearly five times the failure rate of the SOTA agent. This confirms that *DeepSearchQA* tasks are intractable via simple semantic search; they require structured retrieval and multi-step reasoning to decompose queries. Below a certain parameter or compute threshold, agents suffer from trajectory divergence, pursuing entirely incorrect search paths that yield zero relevant results. This finding suggests that scaling down to a cheaper model (like *Claude 4.5 Sonnet* or *Gemini 2.5 Flash*) for research tasks is not a linear trade-off: it is a step-function drop in utility.

Metric Divergence: The Last Mile Problem. A core contribution of *DeepSearchQA* is evaluating the gap between "good" retrieval and "exhaustive" retrieval. We observe a distinct divergence between the granular F1-Score and strict Fully Correct ($S = G$) success rates. For instance, while the *Gemini Deep Research Agent* achieves an F1-Score of 81.90%, its strict success rate is 66.09%. This roughly 15-point gap represents the *"Last Mile Problem"* in autonomous deep research. For *GPT-5 Pro High Reasoning*, the gap is around 13 points. This is driven by two opposing failure modes: *under-retrieval*, where the agent finds the majority of items but misses the "long tail" of obscure entities; and *over-retrieval*, where the agent achieves perfect recall ($R = 1.0$) but fails to recognize when the search is complete, hallucinating extra items or drifting into adjacent topics. This gap validates the necessity of strict set-based metrics for agent evaluation because high recall is insufficient if the agent lacks the discernment to filter noise.

Full Prompt	Failure Mode	Model Response Snippet
Which EU country, that does not touch the Mediterranean Sea, had **the least annual transport** of goods through air, sea, inland waterway, rail, and road by thousand tonnes according to Eurostat in 2022 where every transit method actually had an amount of transported goods. If a country didn't use all 5 methods here, discard that country.	**Quantitative Estimation Error:** *Gemini Deep Research Agent* correctly identifies all countries that satisfy the conditions in the prompt but fails to rank them correctly. It cannot find concrete information and approximates the quantities which creates the error.	"Even a conservative **estimate** places this in the tens of millions ... Total **Estimated** Tonnage".
What are the IATA codes for the US major airports, as classified by the US Department of Transportation, which saw improvements of at least one ranking position of their **On-Time Arrival Rankings** from the 2023 calendar year to the 2024 calendar year, where the on time percentage in 2024 was at least 77.0% and no greater than 81.5%? Use Bureau of Transportation Statistics for data.	**Tool Call Limitation:** *GPT-5 Pro high reasoning* finds an excel file, which supposedly contains all data. However, it states that it can not open it and stops. *Gemini Deep Research Agent* encounters the data as well but continues to search other sources and answers the question partially correct.	"this exactly from ... major airport's ... **rank and on-time arrival** percentage back to 2003 ... only as an Excel download, and **I can't open Excel files directly in this chat.**"
Of all the botanists who have papers in the "papers of individual botanists" section of the Sherardian Library **(MSS. Sherard 26-448)**, which ones never held the position of Sherardian Professor of Botany? Please list their full names in the order that their work appears in the Sherardian Library.	**Stopping Criterion Failure:** *GPT-5 Pro high reasoning* correctly identifies the list of authors on the described work. However, it then does not correctly filter the list according to the constraints from the prompt.	"Sources and notes: ... lists the named individuals **with ...** and presents them in the same **alphabetical** order ..."

Table 5 | Failure modes on DeepSearchQA. *Gemini Deep Research Agent* and *GPT-5 Pro High Reasoning* have common failure patterns. The first column shows prompts that elicit failures; the second column categorizes the specific reasoning or tool limitation; and the third column provides exact text excerpts from the corresponding model demonstrating the error.

5. Future Work

5.1. Limitations

While *DeepSearchQA* offers a robust framework for evaluating comprehensive retrieval, it relies on specific design choices that entail certain limitations. By employing an exclusively outcome-based evaluation, we effectively treat the agent as a black box. In the absence of trajectory data, it is difficult to distinguish between an agent that reasoned correctly and one that arrived at the correct list through inefficient or accidental means (e.g., lucky guessing). Additionally, the *Static Web Assumption*, while necessary for reproducibility, limits the evaluation of "breaking news" retrieval where ground truth is volatile. A task's ground truth may become outdated if source websites are removed or their content is significantly altered. This is a prevalent challenge for all benchmarks operating on the live web, necessitating periodic manual reviews and updates to the dataset.

5.2. Methodological Extensions

The DeepSearchQA benchmark provides a foundation that can be extended in several promising directions:

- **Incorporating Process-Based Metrics:** Future iterations of the benchmark could categorize agent trajectories (e.g., pages visited, query sequences). While scoring would remain

outcome-based, this auxiliary data would provide diagnostic insights into failure modes, helping researchers differentiate between retrieval failures, reasoning errors, and synthesis issues.
- **Dynamic and Time-Sensitive Lists:** To test real-time information retrieval, future versions could introduce "live" questions where the ground truth is volatile (e.g., "List all current members of the UK Parliament's Science and Technology Committee"). This would evaluate an agent's ability to handle temporal constraints and fetch fresh data.
- **Weighted Relevance Scoring:** While the current benchmark treats all ground-truth items with equal weight, future extensions could implement graded relevance. By distinguishing between "core" and "peripheral" answers, the framework can leverage rank-aware metrics such as normalized Discounted Cumulative Gain (nCDG) (Järvelin and Kekäläinen, 2002).

5.3. Implications for Advancements in Agent Architecture

By explicitly rewarding exhaustive retrieval, we expect DeepSearchQA to catalyze research for agents that possess a new set of sophisticated skills. High performance on this benchmark will likely require:

- **Systematic Exploration Strategies:** Agents must move beyond opportunistic keyword searching toward methodical exploration strategies (e.g., tree-based or graph-based navigation) to ensure that no relevant sub-pages or sections are overlooked.
- **Advanced Information Synthesis:** As agents gather candidates from heterogeneous sources, they require robust mechanisms to merge information, resolve entity ambiguities, and de-duplicate semantically equivalent answers to form a clean final list.
- **Dynamic Stopping Criteria:** Perhaps the most critical capability is the ability to reason about search completeness. Agents must develop a *stopping criterion* to dynamically determine when the retrieved set is likely exhaustive and further searching would be unproductive—a significant step forward in agentic reasoning.

6. Conclusion

The rapid advancement of LLM-based agents necessitates a corresponding evolution in the methodologies used to evaluate them. While existing benchmarks like Humanity's Last Exam and BrowseComp have been instrumental in pushing the frontiers of agentic reasoning and information discovery, the *Comprehensiveness Gap* remains a critical blind spot. *DeepSearchQA* addresses this by providing a rigorous, set-based evaluation framework that penalizes both under-retrieval and hedging.

DeepSearchQA directly measures the skills of systematic exploration, multi-source synthesis, and search completion—competencies that are vital for many real-world research and analysis tasks. The benchmark's outcome-based evaluation protocol, grounded in the established information retrieval metrics of Precision, Recall, and F1-score, provides a robust and flexible framework for assessing performance without constraining architectural innovation.

Our evaluation reveals that even the most advanced models, such as GPT-5 Pro and Gemini Deep Research Agent, struggle to balance the trade-off between recall and precision. By creating a clear, measurable target for exhaustive retrieval, *DeepSearchQA* aims to catalyze the next phase of agent development—moving from agents that can answer a question to agents that can master a topic and mapping an information landscape. Optimizing for this deeper level of comprehensiveness is the critical next step toward building genuinely capable and reliable autonomous web agents.

7. Contributions and Acknowledgements

- **Experimental design:** Nikita Gupta, Riju Chatterjee and Hidekazu Oiwa created the experimental design behind the benchmark and ran all the reported experiments.

- **Organization:** Connie Tao, Dipanjan Das, Lukas Haas and John Blitzer managed the overall organization of the effort from start to completion.

- **Early experimentation:** Emily Ingebricson, Chang Liu, Megha Mohabey, Wanzheng Zhu, Oscar Chacaltana, Jan Ackermann, Andrew Wang, Michael Collins and Sasha Goldshtein contributed to ideas, conducted data collection and ran experiments.

- **Sponsors:** Srinivasan (Cheenu) Venkatachary, Koray Kavukcuoglu, Slav Petrov, Ya Xu, and Yossi Matias sponsored the effort and provided technical guidance.

All authors wrote parts of the report.

We would also like to thank:

- **Gemini team** for the support and model access.
- **Kaggle team** for their expertise, releasing the leaderboard, and running final evaluations.
- **Expert data annotators** who helped to collect examples in the paper.
- Our reviewers **Kristina Toutanova**, **Jayant Madhavan**, **Chris Alberti** and **Phoebe Kirk** for valuable feedback.

References

Anthropic. Introducing claude 4, 2025. URL https://www.anthropic.com/news/claude-4.

J. A. Bishop, Q. Xie, and S. Ananiadou. LongDocFACTScore: Evaluating the factuality of long document abstractive summarisation. *arXiv preprint arXiv:2309.12455*, 2023.

S. Chen, Y. Zhao, J. Zhang, I.-C. Chern, S. Gao, P. Liu, and J. He. Felm: Benchmarking factuality evaluation of large language models. *arXiv preprint arXiv:2310.00741*, 2023.

A. Cheng, A. Jacovi, A. Globerson, B. Golan, C. Kwong, C. Alberti, C. Tao, E. Ben-David, G. S. Tomar, L. Haas, Y. Bitton, A. Bloniarz, A. Bai, A. Wang, A. Siddiqui, A. B. Castillo, A. Atias, C. Liu, C. Fry, D. Balle, D. Ghosal, D. Kukliansky, D. Marcus, E. Gribovskaya, E. Ofek, H. Zhuang, I. Laish, J. Ackermann, L. Wang, M. Risdal, M. Barnes, M. Fink, M. Amin, M. Ambar, N. Potikha, N. Gupta, N. Katz, N. Velan, O. Roval, O. Ram, P. Zablotskaia, P. Bang, P. Agrawal, R. Ghiya, S. Ganapathy, S. Baumgartner, S. Erell, S. Prakash, T. Sellam, V. Rao, X. Wang, Y. Akulov, Y. Yang, Z. Yang, Z. Lai, Z. Wu, A. Dragan, A. Hassidim, F. Pereira, S. Petrov, S. Venkatachary, T. Doshi, Y. Matias, S. Goldshtein, and D. Das. The facts leaderboard: A comprehensive benchmark for large language model factuality, 2025. URL https://deepmind.google/blog/facts-benchmark-suite-systematically-evaluating-the-factuality-of-large-language-models/.

DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

M. Du, B. Xu, C. Zhu, X. Wang, and Z. Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *arXiv preprint arXiv:2101.02235*, 2021.

L. Haas, G. Yona, G. D'Antonio, S. Goldshtein, and D. Das. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *arXiv preprint arXiv:2509.07968*, 2025.

O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, and Y. Matias. TRUE: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*, 2022.

C.-W. Huang and Y.-N. Chen. FactAlign: Long-form factuality alignment of large language models. *arXiv preprint arXiv:2410.01691*, 2024.

A. Jacovi, M. Ambar, E. Ben-David, U. Shaham, A. Feder, M. Geva, D. Marcus, and A. Caciularu. Cover-bench: A challenging benchmark for complex claim verification. *arXiv preprint arXiv:2408.03325*, 2024.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611. Association for Computational Linguistics, July 2017.

J. Kasai, K. Sakaguchi, Y. Takahashi, R. L. Bras, A. Asai, X. Yu, D. Radev, N. A. Smith, Y. Choi, and K. Inui. Realtime qa: What's the answer right now? *arXiv preprint arXiv:2207.13332*, 2024.

S. Krishna, K. Krishna, A. Mohananey, S. Schwarcz, A. Stambler, S. Upadhyay, and M. Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*, 2025.

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

H. Li, S. Li, F. Hao, C. J. Zhang, Y. Song, and L. Chen. Booster: Leveraging large language models for enhancing entity resolution. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1043–1046. Association for Computing Machinery, 2024. ISBN 9798400701726.

J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*, 2023.

S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2022.

Y. Liu, N. Moosavi, and C. Lin. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.753.

G. Mialon, C. Fourrier, C. Swift, T. Wolf, Y. LeCun, and T. Scialom. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.

S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100. Association for Computational Linguistics, Dec. 2023. URL https://aclanthology.org/2023.emnlp-main.741/.

M. Miroyan, T.-H. Wu, L. King, T. Li, J. Pan, X. Hu, W.-L. Chiang, A. N. Angelopoulos, T. Darrell, N. Norouzi, and J. E. Gonzalez. Search arena: Analyzing search-augmented llms. *arXiv preprint arXiv:2506.05334*, 2025.

OpenAI. Learning to reason with LLMs, 2024. URL https://openai.com/index/learning-to-reason-with-llms.

OpenAI. Introducing deep research, 2025a. URL https://openai.com/index/introducing-deep-research/.

OpenAI. Gpt-5 system card, 2025b. URL https://openai.com/index/gpt-5-system-card/.

OpenAI. Openai o3 and o4-mini system card, 2025c. URL https://openai.com/index/o3-o4-mini-system-card/.

L. Phan, A. Gatti, Z. Han, N. Li, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

S. Ramprasad and B. C. Wallace. Do automatic factuality metrics measure factuality? A critical evaluation. *arXiv preprint arXiv:2411.16638*, 2024.

H. Rashkin, V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter. Measuring attribution in natural language generation models. *arXiv preprint arXiv:2112.12870*, 2022.

Y. Song, Y. Kim, and M. Iyyer. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. *arXiv preprint arXiv:2406.19276*, 2024.

L. Tang, P. Laban, and G. Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. *arXiv preprint arXiv:2404.10774*, 2024.

T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar, Y.-H. Sung, D. Zhou, Q. Le, and T. Luong. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*, 2023.

J. Wei, N. Karina, H. W. Chung, Y. J. Jiao, S. Papay, A. Glaese, J. Schulman, and W. Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024a.

J. Wei, C. Yang, X. Song, Y. Lu, N. Hu, J. Huang, D. Tran, D. Peng, R. Liu, D. Huang, C. Du, and Q. V. Le. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*, 2024b.

J. Wei, Z. Sun, S. Papay, S. McKinney, J. Han, I. Fulford, H. W. Chung, A. T. Passos, W. Fedus, and A. Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

X. Yang, K. Sun, H. Xin, Y. Sun, N. Bhalla, X. Chen, S. Choudhary, R. D. Gui, Z. W. Jiang, Z. Jiang, L. Kong, B. Moran, J. Wang, Y. E. Xu, A. Yan, C. Yang, E. Yuan, H. Zha, N. Tang, L. Chen, N. Scheffer, Y. Liu, N. Shah, R. Wanga, A. Kumar, W. tau Yih, and X. L. Dong. Crag – comprehensive rag benchmark. *arXiv preprint arXiv:2406.04744*, 2024.

Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

W. Zhang, J. Liao, N. Li, K. Du, and J. Lin. Agentic information retrieval. *arXiv preprint arXiv:2410.09713*, 2024.

W. Zhao, T. Goyal, Y. Y. Chiu, L. Jiang, B. Newman, A. Ravichander, K. Chandu, R. L. Bras, C. Cardie, Y. Deng, et al. WildHallucinations: Evaluating long-form factuality in LLMs with real-world entity queries. *arXiv preprint arXiv:2407.17468*, 2024.

A. Judge Prompt Templates

Grader Prompt

```
Your task is to evaluate whether a given "AI Response" for a specific "User Prompt"
arrived at the correct answer.

**Answer Correctness Task**

*   **Purpose:** Assess whether the AI response provides the correct answer(s) based on
the provided "Correct Answer" and "Prompt Type".
*   **Process:**
    *   Identify the "Prompt Type": "<prompt_type>".
    *   Refer to the "Correct Answer": "<answer>".
    *   Based on the "Prompt Type", determine if the "AI Response" contains the expected
    answer(s).
        *   **'Single Answer'**: Check if the response provides the answer that addresses
        the user's question. It does not have to match the exact wording of the provided
        answer.
        *   **'Set Answer'**: Check if the response includes *each* item from the provided
         ground truth answers. The order might not matter unless specified otherwise. The
         response might include more answers than the list. Determine the correctness *
        only* based on the list first and then check if the response includes answers not
         in the list.
    *   **Explanation:** Provide a brief explanation justifying your assessment of answer
     correctness, referencing specific parts of the AI response and the correct answer.
    *   **Correctness Details:** Provide a dictionary, one key for each expected answer
    part, and value is a boolean indicating whether each expected answer part was found.
        *   For 'Set Answer', this will be a list of attributes, one for each item/part in
         the "Correct Answer". Each key will be a string indicating the expected answer
        part, and the value will be a boolean indicating whether that part was found in
        the response.
    *   **Excessive Answers:** Provide a list of strings, each indicating an excessive
    answer part. If the response provides answers that are **not** in the "Correct Answer
    " list, add these answers as excessive answers. Return an empty list when there's no
    excessive answers in the response.


**Output Format:**

Your evaluation *must* be structured as a nested JSON dictionary with the following top-
level keys: '"Answer Correctness"'. Please return NULL if any of "Prompt", "AI Response"
 or "Correct Answer" is empty.
The value for '"Answer Correctness"' should be a dictionary containing '"Explanation"' (
a string), '"Correctness Details"' (a dictionary where each key is the expected correct
answer, and the value is a boolean indicating whether the response contains the correct
answer), and '"Excessive Answers"' (a list of strings indicating the excessive answers).

Make sure you return a valid JSON string. Pay special attention to quotes, commas and
special characters in the JSON string. Make sure to escape all special characters and
quotes in the JSON string.
```

Grader Partial Output Example

```
**Example (Partial):**

"```json
{{
  "Answer Correctness": {{
    "Explanation": "The response correctly identified Belgium and France but also
    includes an excessive answer, Italy.",
    "Correctness Details": {{
      "Belgium": true,
      "France": true,
    }},
    "Excessive Answers": [ "Italy" ]
  }}
}}
```"

**Now, proceed with the evaluation using the provided User Prompt, AI Response, and
Correct Answer.**

User Prompt (Wrapped in <prompt> and </prompt>):
<prompt>
{prompt}
</prompt>

** Correct Answer (Wrapped in <answer> and </answer>):
Prompt Type: {prompt_type}
<answer>
{answer}
</answer>

AI assistant response (Wrapped in <response> and </response>):
<response>
{response}
</response>

Rating:'''
```