

ResearcherBench: Evaluating Deep AI Research Systems on Open-ended AI Research Tasks

Tianze Xu^{1,3*} Pengrui Lu^{1,2,3*} Lyumanshan Ye^{1,3} Xiangkun Hu² Pengfei Liu^{1,2,3†}

¹Shanghai Jiao Tong University ²Shanghai Innovation Institute ³GAIR
{tzxu_1213, lupengrui, yelyumanshan, pengfei}@sjtu.edu.cn, xkhu17@fudan.edu.cn

Abstract

Deep research agents have emerged with capabilities extending from basic queries to sophisticated research tasks. However, existing benchmarks primarily evaluate their abilities on web retrieval and report generation, overlooking the potential for discovering, integrating and generating insights in AI research. To address this gap, we introduce ResearcherBench, the first benchmark focused on evaluating the capabilities of Deep AI Research Systems (DARS) on open-ended AI research tasks. We curated a dataset of 65 questions expertly selected from real-world AI research scenarios, spanning 35 different AI subjects and categorized into 3 different types. Our dual evaluation framework consists of two components: insight assessment which implements expert-designed rubrics to evaluate insight quality, and citation assessment measuring citation faithfulness and groundedness (citation density). We evaluated several leading commercial DARS and baseline systems. Our evaluation results reveal that these systems excel at open-ended consulting questions compared to technical implementation questions, demonstrating the potential for DARS to serve as AI research partners. We open-source ResearcherBench and evaluation framework to accelerate the development of AI research assistants capable of scientific collaboration: <https://github.com/GAIR-NLP/ResearcherBench>.

1 Introduction

The advent of artificial intelligence has fundamentally transformed complex problem-solving, with deep research systems emerging as AI agents capable of autonomously conducting research workflows (Gridach et al., 2025; Huang et al., 2025). These systems demonstrate proficiency in information retrieval, analysis, and synthesis across diverse domains (Xu & Peng, 2025), with increasing deployment in AI research for technical investigation, literature review, and knowledge synthesis (Stanford HAI, 2025).

These applications have demonstrated value in research workflows and enhancing productivity in well-established AI research domains. However, the community might overlook an important capability: their potential to assist researchers in exploring open-ended AI research tasks at the cutting edge of AI knowledge (Lu et al., 2024; Research, 2025).

The transition from systems that merely retrieve and summarize information to research partners capable of collaboration in open-ended AI research represents a major challenge in developing AI research assistants today (Wang & Chen, 2024). We define this emerging category as **Deep AI Research Systems (DARS)**: agentic systems that combine dynamic reasoning with autonomous research workflow execution, including multi-iteration web retrieval, tool utilization, and iterative refinement processes (Huang et al., 2025).

This capability gap raises a critical question: can current DARS truly assist human researchers in tackling the challenging, high-valued, and open-ended AI research questions,

*Equal contribution.

†Corresponding author.

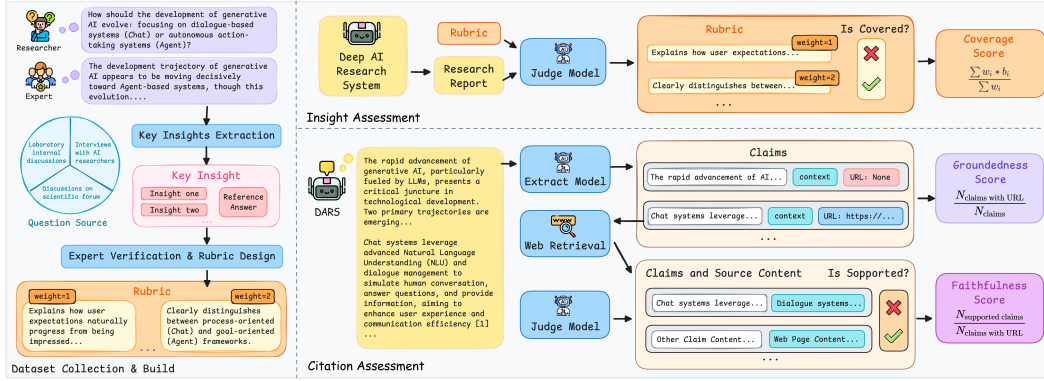


Figure 1: ResearcherBench Framework Overview. The framework consists of three main components: (1) Dataset collection from authentic research scenarios leading to expert-generated rubrics, (2) Insight assessment to evaluate coverage against rubrics, and (3) Citation assessment to measure faithfulness and groundedness scores.

where definitive answers do not yet exist and novel insights must be synthesized from fragmentary and cross-domain information within the artificial intelligence field?

Answering this question demands novel evaluation frameworks. Existing benchmarks for evaluating DARS capabilities predominantly focus on comprehensive report generation abilities (Du et al., 2025) or web interaction capabilities (Gou et al., 2025; Wei et al., 2025), failing to assess the ability to understand, analyze, and provide meaningful insights on highly specialized, open-ended AI research problems (Zheng et al., 2025; Starace et al., 2025).

To address this gap, we introduce **ResearcherBench**, a benchmark specifically designed to evaluate DARS on open-ended AI research tasks, as shown in Figure 1. We initially collected an initial corpus of 932 questions from authentic AI research scenarios, and filtered them to a curated dataset of 65 high-quality questions spanning 35 different AI subjects and categorized into three types. We also developed a dual evaluation framework combining human expert-designed insight assessment with citation assessment, to evaluate both the intellectual depth and factual reliability of DARS responses.

ResearcherBench complements existing evaluation methods by focusing on whether DARS can provide meaningful assistance to human researchers working on unsolved, open-ended problems in the field of AI. Our benchmark represents a shift in evaluation focus — moving from assessing “whether systems can retrieve and summarize information”, to evaluating “whether DARS can understand complex problems and provide meaningful insights as effective research assistants.” Our main contributions are stated as follows:

- **A Novel Task Collection:** We present an expert-curated dataset of 65 high-quality questions sourced from authentic AI research scenarios, spanning 35 distinct AI subjects and categorized into three types.
- **A Dual Evaluation Framework:** Our assessment methodology combines human expert-designed insight assessment with citation assessment, evaluating both insight quality and citation reliability.
- **Comprehensive Empirical Analysis:** Our evaluation of commercial DARS reveals their superior performance on open-ended consulting over technical synthesis questions, and a “high faithfulness, low groundedness” pattern in citation assessment.
- **Open-Source Contribution:** We release ResearcherBench as a comprehensive evaluation platform, encompassing our curated dataset and evaluation framework to collaboratively foster the development of AI research assistance capabilities.

Benchmark	Requires Real-time Web	Open-ended Research Query	Human Expert- designed Rubric
ResearchBench (Liu et al., 2025)	✗	✓	✗ (Automated)
Mind2Web 2 (Gou et al., 2025)	✓	✗	✗ (LLM-generated)
DeepResearch Bench (Du et al., 2025)	✓	✗	✗ (LLM-generated)
ReportBench (Li et al., 2025b)	✓	✓	✗ (Reference-based)
SciArena (Zhao et al., 2025)	✗	✓	✗ (Human Vote)
ResearchQA (Yifei et al., 2025)	✗	✓	✗ (Survey-mined)
ResearcherBench (Ours)	✓	✓	✓

Table 1: Comparison of Deep Research-related Benchmarks

2 Related Work

2.1 Deep AI Research Systems

Early DARS. Early Deep AI Research Systems (DARS) evolved from basic prompt-based approaches (Zheng et al., 2024; Alzubi et al., 2025) and supervised fine-tuning methods (Asai et al., 2024), built upon retrieval-augmented generation (RAG) architectures within static knowledge repositories. These systems faced limitations including reliance on non-updatable knowledge bases and lack of iterative reasoning for complex research workflows.

Modern DARS. Modern DARS address these limitations through reinforcement learning-based approaches (Zheng et al., 2025; Jin et al., 2025; Chen et al., 2025; Song et al., 2025), implementing multi-agent architectures for autonomous web search with iterative query formulation and dynamic refinement. Commercial implementations of Deep Research System (OpenAI, 2025b; Google, 2025; xAI, 2025; Anthropic, 2025) demonstrate these capabilities, generating high-quality research reports through iterative workflows.

2.2 Research-related Evaluation Benchmarks

Traditional RAG-based Benchmarks. Early evaluation frameworks established multi-hop reasoning assessment requiring information synthesis across sources (Joshi et al., 2017; Yang et al., 2018), while OpenScholar (Asai et al., 2024) emphasized the citation accuracy of literature synthesis with expert-written responses. These benchmarks focused on factual retrieval and basic reasoning.

Specialized Research Task Benchmarks. ResearchBench (Liu et al., 2025) evaluates scientific hypothesis discovery, while DeepResearch Bench (Du et al., 2025) introduces RACE and FACT to assess report generation and information retrieval. PaperBench (Starace et al., 2025) and EXP-Bench (Kon et al., 2025) evaluate research replication on academic tasks. These benchmarks focus on established technical competencies and knowledge synthesis.

Academic Report and Scholarly QA Benchmarks. ReportBench (Li et al., 2025b) scores generated surveys against paper references for citation quality and faithfulness, SciArena (Zhao et al., 2025) collects human preference votes on literature-grounded answers, and ResearchQA (Yifei et al., 2025) mines 21K queries and rubric items from surveys across 75 fields. These evaluation signals come from survey-mined rubrics or human votes.

Agentic Search and Browsing Benchmarks. Mind2Web 2 (Gou et al., 2025) introduces 130 long-horizon real-time browsing tasks with an Agent-as-a-Judge evaluation framework, while BrowseComp (Wei et al., 2025) measures persistent browsing through navigation tasks. Both prioritize information retrieval breadth and web interaction.

Limitations. Table 1 presents a comparison of several related benchmarks. Existing benchmarks focus on competencies like report generation and web interaction, prioritizing information retrieval over the conceptual understanding and insight generation, which are required for open-ended AI research. Moreover, evaluation rubrics for AI research questions cannot be reliably generated by automated systems, requiring extensive AI domain expertise to assess valuable insights, as discussed in Section A.

This gap motivates the need for evaluation frameworks that can assess AI systems’ capabilities as research assistants in exploring open-ended AI research tasks, where insight generation and deep understanding of AI problems are important.

3 ResearcherBench

ResearcherBench presents a systematic approach to constructing a comprehensive benchmark for evaluating DARS capabilities on open-ended AI research tasks. We employed rigorous data collection and filtering methodologies to curate AI research questions from real-world scientific scenarios, resulting in a high-quality dataset of 65 questions across 35 subjects, as shown in Figure 2.

3.1 Data Collection and Categorization

We collected AI research questions from authentic scientific scenarios, resulting in an initial corpus of 932 candidate questions. These questions are from three primary sources: laboratory discussions, interviews with AI researchers, and scientific forum discussions. Questions were deemed “open-ended AI research questions” if they target core, unsolved limitations of current AI domains, requiring novel insights to address scientific challenges (Marcus, 2020). Definitions and examples for each question type are provided in Appendices C.1.

We classified these questions into three types based on their cognitive demands: Technical Details, Literature Review, and Open Consulting. The subject-based classification yielded 35 distinct AI research areas including model architecture, multimodal fusion, AI ethics, and other emerging paradigms, ensuring comprehensive coverage of contemporary AI research challenges. Details are provided in Appendix C.2.

3.2 Question Selection and Filtering

We developed a type-specific selection framework to filter high-quality AI research questions. Instead of uniform selection standards, we designed tailored criteria for each question type: for example, technical details questions emphasized methodological precision and verifiability, literature review questions focused on comprehensive scope and consistency, while open consulting questions prioritized conceptual depth and innovative perspectives.

Each question received independent evaluation from at least two experienced annotators with domain expertise in the corresponding AI field. Through this systematic review process, we refined the initial corpus from 932 to 65 high-quality questions that met our stringent selection standards. This selective framework prioritizes quality over dataset size, aligning with our goal of assessing DARS capabilities on genuinely open-ended and challenging AI research scenarios. Detailed evaluation methodology, criteria and examples are provided in Appendices C.3 and C.4.

4 Dual Evaluation Framework

ResearcherBench introduces a dual evaluation framework that comprehensively assesses DARS performance through both **insight assessment** (evaluating insight quality and conceptual depth) and **citation assessment** (measuring citation accuracy and content groundedness). This complementary approach captures both the intellectual contribution and empirical reliability of research responses.

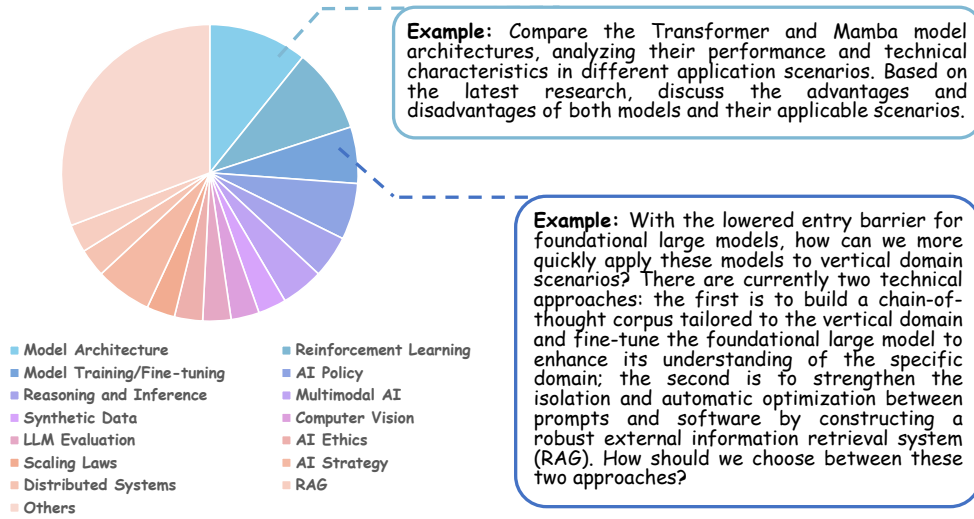


Figure 2: ResearcherBench topic distribution with representative examples. Left: pie chart showing the distribution of the 35 AI subjects in the benchmark. Right: concrete question examples from major subjects.

4.1 Insight Assessment

Traditional LLM-as-a-Judge evaluation (Gu et al., 2025; Li et al., 2025a) with single overall ratings provides insufficient granularity for assessing complex research responses. Open-ended AI research evaluation requires nuanced assessment across multiple dimensions that scores cannot capture.

Our fine-grained insight assessment framework addresses this limitation by decomposing complex research questions into multiple specific, expert-designed criteria with assigned importance weights. This approach enables systematic evaluation of research insight quality across conceptual understanding, methodological rigor, and analytical depth. This structured approach provides the granular assessment necessary for evaluating the quality of complex research insights.

4.1.1 Rubric Construction

The evaluation rubrics are designed by experts in various AI domains to reflect the standards and expectations of open-ended AI research. We employ a three-stage method to construct these rubrics:

Stage 1: Insight Extraction. For each question, we collect and analyze a diverse range of contextual sources for each question, including original discussion records, expert opinions, and technical background materials. From these multifaceted inputs, we distill key insights to construct reference materials. These materials then undergo expert review and supplementation, to establish a robust foundation for subsequent rubric development.

Stage 2: Criteria Design. Following standardized rubric design guidelines, we invited human annotators (experienced masters, Ph.D.s or professionals specialized in AI) to design evaluation rubrics with multiple criteria independently. The reference materials constructed in the Stage 1 serve as auxiliary information. Each criterion addresses a specific aspect of reasoning or insight, assigned an importance weight (1–3) based on its impact on overall answer quality. Guidelines for rubric construction and examples are provided in Appendix D.

Stage 3: Quality Control. To ensure rubric reliability and validity, we implemented a multi-stage quality control process. Each rubric was collaboratively developed by 2-3 experienced AI researchers, with one annotator responsible for initial drafting and weight assignment, and the others conducting comprehensive review and refinement. All rubrics were assessed for completeness, clarity, independence, and discriminative power. We conducted pilot testing with 20 DARS responses to identify and refine problematic criteria through iterative revision. Finally we validated rubric effectiveness through meta-evaluation, comparing expert judgments with automated assessments, as detailed in Section 4.3.

The resulting benchmark comprises a total of 931 expert-designed rubric criteria across 65 questions. Table 2 summarizes their key distributional properties. Further breakdowns of the criteria-count and weight distributions are provided in Appendix E.

Metric	Min	Max	Mean
Criteria count per question	6	21	14.32
Total score per question	14	39	25.52

Table 2: Overall rubric statistics across all 65 benchmark questions.

4.1.2 Evaluation Metric

We define Coverage Score S_c as a weighted metric of how comprehensively a DARS response cover the key insights specified in the expert-designed rubric. Let Q_k denote the k -th question in our benchmark, and $\text{Res} = \text{DARS}(Q_k)$ represent the final response generated by DARS. The binary indicator $b_i \in \{0, 1\}$ for each criterion is computed as:

$$b_i = \text{Judge}(Q_k, \text{Res}, r_i), \quad (1)$$

where r_i represents the content of the i -th criterion, and $\text{Judge}(\cdot)$ returns 1 if the response satisfies criterion r_i , 0 otherwise. For the question Q_k , the Coverage Score is calculated as:

$$S_c = \frac{\sum_{i=1}^n w_i \cdot b_i}{\sum_{i=1}^n w_i}, \quad (2)$$

where $w_i \in \{1, 2, 3\}$ represents the weight of the i -th criterion, and n is the total number of criteria. This formulation reflects both the coverage of expert-aligned criteria and the relative importance across different criteria.

4.2 Citation Assessment

While insight assessment evaluates insight quality and conceptual depth, factual reliability remains a fundamental requirement for DARS-generated research reports (Zhang et al., 2023). Our citation assessment framework adopts two metrics: *faithfulness* and *groundedness* (ExplodingGradients, 2024) in automated research synthesis.

4.2.1 Assessment Methodology

Stage 1: Claim Extraction. We employ an extraction model to identify all factual claims in DARS-generated reports with their contexts. For each extracted claim, the model determines whether it corresponds to a citation URL in the report’s reference section. Claims with valid citations are saved as URL-claim-context triplets for verification. Otherwise, this claim is considered ungrounded, and its URL is marked as empty.

For question Q_k , let $\text{Res}, \text{Ref} = \text{DARS}(Q_k)$ represent the main content and reference section of the research report generated by DARS. We define a claim c_i as a triplet $(\text{text}_i, \text{context}_i, \text{url}_i)$, where text_i is the textual content of the claim extracted from responses, $\text{url}_i \in \text{URL} \cup \{\emptyset\}$ is the associated citation URL extracted from references (if exists), and context_i is the context surrounding the claim used as supplementary information for verification. The complete set of extracted claims C_k for question Q_k is then represented as:

$$C_k = \{c_i \mid c_i = \text{Extract}(Q_k, \text{Res}, \text{Ref}), i = 1, 2, \dots, N_k\}, \quad (3)$$

where $\text{Extract}(Q_k, \text{Res}, \text{Ref})$ is the extraction model that identifies factual statements in the Res, and $N_k = |C_k|$ is the total number of claims for Q_k .

Stage 2: Citation Support Verification. For each claim, we extract the entire textual contents from URL sources using the Jina Reader API¹. Then a judge model performs binary verification to determine whether the extracted content supports the corresponding claim. When claims are semantically incomplete or ambiguous in isolation, the surrounding context serves as supplementary information to assist the model’s judgment.

Denoted that $C_k^{\text{cited}} \subseteq C_k$ is the subset of cited claims with URLs, and $C_k^{\text{supp}} \subseteq C_k^{\text{cited}}$ is the subset of claims that are both cited and supported by their URL sources, we define:

$$\begin{aligned} C_k^{\text{cited}} &= \{c_i = (\text{text}_i, \text{context}_i, \text{url}_i) \mid c_i \in C_k \text{ and } \text{url}_i \neq \emptyset\}, \\ C_k^{\text{supp}} &= \{c_i \in C_k^{\text{cited}} \mid \text{Judge}(\text{text}_i, \text{context}_i, \text{Text}(\text{url}_i)) = 1\}, \end{aligned}$$

where $\text{Text}(\text{url})$ represents the entire textual content extracted from the URL, and $\text{Judge}(\text{text}, \text{context}, \text{Text}(\text{url}))$ returns 1 if the claim is supported by the URL, 0 otherwise.

4.2.2 Evaluation Metric

Based on verification results, two complementary metrics are calculated to assess the overall factual reliability of DARS-generated report:

- **Faithfulness Score S_f** , which measures the accuracy of citations in supporting their corresponding claims. This metric evaluates the proportion of cited claims that are actually supported by their referenced sources, indicating the reliability of citation-claim relationships when citations are provided.
- **Groundedness Score S_g** , which characterizes the citation density of a response. As a descriptive indicator, this metric quantifies the proportion of factual claims that carry explicit citation support: high-groundedness responses lean toward source compilation, while low-groundedness responses lean toward cross-source synthesis.

Suppose that $N_{c,k} = |C_k^{\text{cited}}|$ is the total number of cited claims, and $N_{s,k} = |C_k^{\text{supp}}|$ is the total number of supported claims, these two metrics are calculated as:

$$S_f = \frac{N_{s,k}}{N_{c,k}}, \quad S_g = \frac{N_{c,k}}{N_k}. \quad (4)$$

4.3 Judge Model Selection Via Human Evaluation

Selecting an appropriate automated judge model is critical for benchmark evaluation. We address this through systematic human expert evaluation: by comparing multiple candidate LLM judges against ground-truth human expert assessments, we can identify which automated judge achieves optimal alignment with expert-level evaluation standards.

This selection process ensures that our automated scoring mechanism can reliably reflect human expert judgment, while enabling comprehensive and scalable evaluation across our full benchmark.

4.3.1 Experimental Design

Human Expert Annotation for Ground Truth. To establish reliable reference standards for judge model selection, we firstly conducted human expert evaluation. Given the intensive nature of expert evaluation (1.5 hours per response for insight assessment), we carefully constructed a stratified validation subset of 20 responses spanning different question types, difficulty levels, and DARS systems to ensure representativeness. We recruited a team of 5 domain expert annotators to independently evaluate these responses using the same

¹ <https://jina.ai/reader>

criteria employed in rubrics assessment. Each response received evaluation from at least two independent annotators to ensure reliability. This process yielded reliable reference standards for judge model selection.

Candidate Judge Models and Selection Criteria. We systematically evaluated five leading LLMs as candidate automated judges: DeepSeek R1, Gemini-2.5-flash, GPT-4.1, o3, and o3-mini. Our selection methodology considered two complementary dimensions: (1) **Performance consistency**: measured by accuracy, precision, recall, and F1-score quantifying agreement between automated judgments and human expert annotations; and (2) **Cost efficiency**: API costs per question, enabling practical large-scale deployment. This dual-criteria approach ensures we select judge models that achieve both high-quality human alignment and practical scalability for comprehensive benchmark evaluation.

4.3.2 Validation Results and Final Selection

Experiment shows that our Inter-annotator agreement (Cohen’s Kappa) achieved $\kappa = 0.67$, which indicates substantial consensus among independent experts. As shown in Table 3, top-performing models (o3, o3-mini, GPT-4.1) achieved F1-scores of 0.80-0.81, demonstrating substantial agreement with human experts (Landis & Koch, 1977).

Based on the results, we selected o3-mini for insight assessment, due to the optimal trade-off between human alignment (F1: 0.80, precision: 0.85) and cost efficiency (\$0.13 per question). For citation assessment, we chose GPT-4.1 for its superior performance in long-context required for claim extraction and citation verification (OpenAI, 2025a). A cross-vendor sensitivity study further shows that three judges from distinct model families (o3-mini, Gemini-2.5-Flash, and DeepSeek R1) yield identical system rankings, confirming that our results are not an artifact of judge choice (Appendix M).

Judge Model	Acc.	Prec.	Rec.	F1	Cost
DeepSeek R1	0.71	0.83	0.68	0.75	0.23
Gemini 2.5 Flash	0.72	0.75	0.82	0.78	0.54
GPT-4.1	0.72	0.75	0.85	0.80	0.19
o3	0.76	0.80	0.83	0.81	0.22
o3-mini	0.76	0.85	0.76	0.80	0.13

Table 3: Performance comparison of candidate judge models against human annotations. Cost represents average API expense per question evaluation (\$). Bold values indicate best performance.

5 Experiments and Results

5.1 Experimental Setup

Evaluated Systems. We evaluated several leading commercial DARS to assess their performance on open-ended AI research tasks: OpenAI Deep Research (OpenAI, 2025b), Gemini Deep Research powered by Gemini-2.5-Pro (Google, 2025), Claude Research (Anthropic, 2025), Grok3 DeepSearch & DeeperSearch (xAI, 2025), and Perplexity Deep Research (Perplexity AI, 2025). We also tested two commercial LLM platforms with Deep Research Mode: Doubao² and Mita³. To provide comprehensive comparison, we evaluated two LLM with web search tools: GPT-4o Search Preview and Perplexity Sonar Reasoning Pro. Details on our specific interaction procedures with each system can be found in Appendix F.

Evaluation Configuration. After comprehensive evaluation of candidate Judge LLMs (detailed in Section 4.3), we selected o3-mini as the judge model for insight assessment to evaluate criteria coverage. For citation assessment, we chose GPT-4.1 as both the extraction and judge model. To mitigate stochastic variation in LLM-based judgment, each criterion is evaluated through three independent inference and the averaged scores are reported. All evaluations were conducted between March and July in 2025 to ensure temporal consistency and fair comparison across systems. Details on data collection, specific implementation (e.g. prompt template), and error analysis are provided in Appendix G.

² <https://www.doubao.com>

³ <https://metaso.cn>

Model	Insight Assessment	Citation Assessment	
	Coverage	Faithfulness	Groundedness
<i>Deep AI Research Systems</i>			
OpenAI Deep Research	70.32% , [-5.0%, +4.9%]	84%, [-4%, +3%]	34%, [-4%, +4%]
Gemini Deep Research	69.29%, [-5.3%, +5.2%]	86% , [-5%, +4%]	59%, [-4%, +3%]
Claude Research [†]	61.13%, [-5.3%, +5.3%]	-	-
Mita Deep Research	58.35%, [-5.7%, +5.9%]	65%, [-4%, +4%]	59%, [-4%, +4%]
Doubao Deep Research	57.54%, [-5.7%, +5.5%]	78%, [-4%, +3%]	63%, [-4%, +4%]
Perplexity Deep Research	48.00%, [-5.6%, +5.5%]	85%, [-4%, +4%]	56%, [-4%, +4%]
Grok3 DeepSearch	44.14%, [-5.4%, +5.7%]	69%, [-6%, +6%]	32%, [-4%, +4%]
Grok3 DeeperSearch	43.98%, [-5.6%, +5.6%]	80%, [-4%, +4%]	31%, [-4%, +4%]
<i>LLM with Search Tools</i>			
GPT-4o Search Preview	35.76%, [-5.1%, +5.2%]	86% , [-4%, +3%]	39%, [-5%, +5%]
Perplexity: Sonar Reasoning Pro	46.63%, [-5.6%, +5.6%]	62%, [-6%, +6%]	68% , [-4%, +4%]

Table 4: Comprehensive evaluation results across Insight Assessment and Citation Assessment. Bold values indicate the best performance for each metric. We report 95% bootstrap confidence intervals (10,000 bootstrap samples) for Coverage, Faithfulness, and Groundedness. [†] Citation assessment for Claude Research is unavailable due to limited citation extraction.

5.2 Main Results

5.2.1 Insight Assessment Results

As shown in Table 4, the performance of DARS fall into three tiers: OpenAI Deep Research and Gemini Deep Research achieve the highest coverage scores (0.70 and 0.69). Claude Research, Mita Deep Research, and Doubao Deep Research demonstrate mid-tier performance (0.58-0.61), while Perplexity Deep Research (0.48) and both Grok variants (0.44) achieving limited coverage. LLM with Search tools generally underperform dedicated DARS across all evaluated metrics, while Perplexity Sonar Reasoning Pro (0.47) outperforms two Grok Deep Research variants (0.44). The bootstrap analysis validates the robustness of our rankings, confirming that the rubric effectively distinguishes between different tiers.

Insight assessment results reveal that: **Iterative web search combined with deep reasoning capability determines high quality in the insight assessment.** Dedicated DARS demonstrate overall superior performance compared to LLM+Search tools (e.g., OpenAI Deep Research v.s. GPT-4o Search Preview), indicating that such models lack ReAct mechanisms (Yao et al., 2022) and deep reasoning capacities, thus failing to address open-ended AI research tasks. Moreover, while Perplexity Sonar Reasoning Pro outperforms Grok DeepSearch despite being a non-DARS system, it remains inferior to other DARS like Perplexity Deep Research, demonstrating that multi-turn ReAct mechanisms (Yao et al., 2022) is essential for comprehensive research synthesis.

5.2.2 Citation Assessment Results

The factual evaluation reveals a consistent “high faithfulness, low groundedness” pattern across the systems. As shown in Table 4, most DARS achieve faithfulness scores of 65%–86%, indicating robust citation-claim relationships, while groundedness scores vary substantially (31%–68%), reflecting different citation densities across systems. Despite being non-DARS systems, GPT-4o Search Preview reaches the highest faithfulness score, and Perplexity Sonar Reasoning Pro demonstrates the highest overall groundedness (68%).

Citation assessment results reveal that: **Different DARS exhibit distinct source-attribution styles, reflecting the diversity of how these systems retrieve and integrate information.** The substantial variation in groundedness scores points to fundamentally different strategies for handling retrieved content: high-groundedness systems behave as “information aggregators” that directly cite and summarize sources, while low-groundedness systems

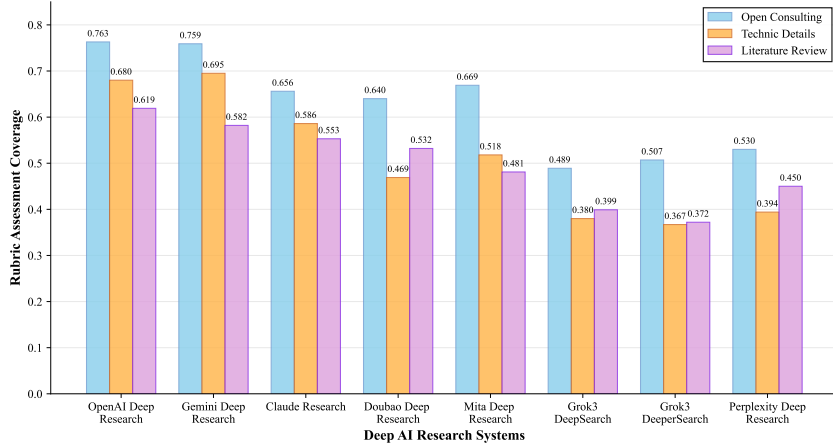


Figure 3: Performance comparison across different question types.

behave as “analytical synthesizers” that transform retrieved information through reasoning. This diversity distinguishes systems which are optimized for source compilation from those optimized for cross-source synthesis and insight generation. Appendix H illustrates how these different attribution styles produce distinct research values.

5.2.3 Analysis across question types

Analysis across question types reveals distinct capability patterns, as illustrated in Figure 3. All systems perform better on open-ended consulting than literature review and technical detail questions, with top performance achieving coverage scores above 76%. We also conduct leave-one-category-out analysis to confirm that the imbalanced category of datasets does not affect system-level rankings in Appendix K.

Results reveal that **Architectural design of DARS favors divergent exploration over precision retrieval**. Open consulting questions benefit from the divergent exploration and creative synthesis at which DARS excel, generating novel insights through broad information integration. On the other hand, Literature review questions require precise retrieval of articles and comprehensive coverage of established work, while technical detail questions demand exact implementation specifics. These differences reveal that current DARS are more effective in exploratory synthesis, positioning them as research ideation tools rather than comprehensive literature surveyors or technical implementation guides.

5.2.4 Score Reliability Analysis

We additionally conduct a score reliability analysis via paired bootstrap significance tests and a repeated-generation experiment, which confirms that the observed performance tiers are statistically significant and stable across generation runs rather than artifacts of noise. Detailed analysis is provided in Appendix J.

6 Conclusion

This paper introduces ResearcherBench, the first benchmark for evaluating Deep AI Research Systems (DARS) on open-ended AI research tasks, comprising 65 curated questions with dual insight and citation assessment frameworks. Our comprehensive evaluation of leading commercial DARS establishes crucial performance baselines and reveals key findings into automated research assistance capabilities. By open-sourcing ResearcherBench, we aim to advance AI evaluation from information retrieval assessment toward research assistance capability, providing the foundation for developing AI systems capable of meaningful AI Research collaboration. Limitations and future directions are discussed in Appendix B.

Acknowledgments

We sincerely thank all the human annotators whose expertise effort were indispensable to the construction and validation of ResearcherBench. Specifically, we are grateful to Yuxiang Zheng, Yixiu Liu, and Yunze Wu for their foundational work on question selection and rubric construction. We also thank Dayuan Fu, Xiaojie Cai, Chenyang Zhou, and Weiye Si for their contributions to meta evaluation and comparative analysis of human-designed v.s. LLM-generated rubrics. Their independent assessments ensured the robustness and consistency of our evaluation framework throughout every stage of the benchmark development.

References

- Salaheddin Alzubi, Creston Brooks, Purva Chiniya, Edoardo Contente, Chiara von Gerlach, Lucas Irwin, Yihan Jiang, Arda Kaz, Windsor Nguyen, Sewoong Oh, et al. Open deep search: Democratizing search with open-source reasoning agents. *arXiv preprint arXiv:2503.20201*, 2025.
- Anthropic. Claude takes research to new places. <https://www.anthropic.com/news/research>, April 2025. Accessed: July 30, 2025.
- Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D’arcy, et al. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*, 2024.
- Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–20, 2025. ISSN 2326-3865. doi: 10.1109/tkde.2025.3554028. URL <http://dx.doi.org/10.1109/TKDE.2025.3554028>.
- Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Fan Yang, Zenan Zhou, Weipeng Chen, Haofen Wang, Jeff Z Pan, et al. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL <https://arxiv.org/abs/2403.04132>.
- Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepre-search bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2025. URL <https://arxiv.org/abs/2404.04475>.
- ExplodingGradients. Ragas: Supercharge your llm application evaluations. <https://github.com/explodinggradients/ragas>, 2024.
- William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022. URL <https://arxiv.org/abs/2101.03961>.
- Google. Gemini deep research - your personal research assistant, 2025. URL <https://gemini.google/overview/deep-research/>. Accessed: April 14, 2025.
- Boyue Gou, Zanning Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanav, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, et al. Mind2web 2: Evaluating agentic search with agent-as-a-judge. *arXiv preprint arXiv:2506.21506*, 2025.
- Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions, 2025. URL <https://arxiv.org/abs/2503.08979>.

- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL <https://arxiv.org/abs/2411.15594>.
- Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, Kun Shao, and Jun Wang. Deep research agents: A systematic examination and roadmap, 2025. URL <https://arxiv.org/abs/2506.18096>.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Serkan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning, 2025. URL <https://arxiv.org/abs/2503.09516>.
- Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396, 2015.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.
- Nina Kirakosyan. Mixture of experts llms: Key concepts explained. <https://neptune.ai/blog/mixture-of-experts-llms>, Jan 2025. Accessed: 2025-09-25.
- Patrick Tser Jern Kon, Jiachen Liu, Xinyi Zhu, Qiuyi Ding, Jingjia Peng, Jiarong Xing, Yibo Huang, Yiming Qiu, Jayanth Srinivasa, Myungjin Lee, et al. Exp-bench: Can ai conduct ai research experiments? *arXiv preprint arXiv:2505.24785*, 2025.
- J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge, 2025a. URL <https://arxiv.org/abs/2411.16594>.
- Minghao Li, Ying Zeng, Zhihao Cheng, Cong Ma, and Kai Jia. ReportBench: Evaluating deep research agents via academic survey tasks. *arXiv preprint arXiv:2508.15804*, 2025b.
- Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. *arXiv preprint arXiv:2503.21248*, 2025.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL <https://arxiv.org/abs/2408.06292>.
- Gary Marcus. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.
- OpenAI. Introducing GPT-4.1 in the API, 2025a. URL <https://openai.com/index/gpt-4-1/>. Accessed: April 14, 2025.
- OpenAI. Introducing deep research, 2025b. URL <https://openai.com/index/introducing-deep-research/>. Accessed: April 14, 2025.
- Perplexity AI. Introducing perplexity deep research, 2025. URL <https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research>. Accessed: April 14, 2025.
- Google Research. Accelerating scientific breakthroughs with an ai co-scientist. <https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/>, 2025.

- Omar Sanseviero, Lewis Tunstall, Philipp Schmid, Sourab Mangrulkar, Younes Belkada, and Pedro Cuenca. Mixture of experts explained, 2023. URL <https://huggingface.co/blog/moe>.
- Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, et al. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *arXiv preprint arXiv:2511.07685*, 2025.
- Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025.
- Stanford HAI. The 2025 ai index report, 2025. URL <https://hai.stanford.edu/ai-index/2025-ai-index-report>. Accessed: April 14, 2025.
- Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: Evaluating ai’s ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.
- Jiayu Wang, Yifei Ming, Riya Dulepet, Qinglin Chen, Austin Xu, Zixuan Ke, Frederic Sala, Aws Albarghouthi, Caiming Xiong, and Shafiq Joty. Liveresearchbench: A live benchmark for user-centric deep research in the wild, 2025. URL <https://arxiv.org/abs/2510.14240>.
- Xiaomei Wang and Xiaoyu Chen. Towards human-ai mutual learning: A new research paradigm, 2024. URL <https://arxiv.org/abs/2405.04687>.
- Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.
- Cameron R. Wolfe. How do mixture-of-experts layers affect transformer models? <https://stackoverflow.blog/2024/04/04/how-do-mixture-of-experts-layers-affect-transformer-models/>, Apr 2024. Accessed: 2025-09-25.
- xAI. Grok 3 beta — the age of reasoning agents, 2025. URL <https://x.ai/news/grok-3>. Accessed: April 14, 2025.
- Renjun Xu and Jingwen Peng. A comprehensive survey of deep research: Systems, methodologies, and applications, 2025. URL <https://arxiv.org/abs/2506.12594>.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.
- Li S. Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. ResearchQA: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics. *arXiv preprint arXiv:2509.00496*, 2025.
- Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren’s song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.
- Yilun Zhao et al. SciArena: An open evaluation platform for foundation models in scientific literature tasks. *arXiv preprint arXiv:2507.01001*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL <https://arxiv.org/abs/2306.05685>.

Yuxiang Zheng, Shichao Sun, Lin Qiu, Dongyu Ru, Cheng Jiayang, Xuefeng Li, Jifan Lin, Binjie Wang, Yun Luo, Renjie Pan, et al. Openresearcher: Unleashing ai for accelerated scientific research. *arXiv preprint arXiv:2408.06941*, 2024.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*, 2025.

A Comparison with Alternative Approaches

To ensure the rigor and reliability of ResearcherBench, we conducted an extensive comparison of our evaluation framework against two widely-adopted alternatives: LLM-generated rubrics and pairwise comparison.

Expert-Designed vs. LLM-Generated Rubrics Prior works employ LLM-generated rubric to construct their benchmarks for its scalability and automation (Du et al., 2025; Gou et al., 2025; Wang et al., 2025). However, this approach faces severe challenges in open-ended research questions: (1) **Circular dependency risk**: models are evaluated based on criteria generated by similar language models, which potentially introduces evaluation biases that favor outputs imitating LLM-typical patterns, rather than genuinely high-quality research; (2) **Lack of domain expertise**: LLMs often fail to identify the truly important aspects of open-ended questions in AI research tasks, instead falling back on generic or easy-to-satisfy criteria; (3) **Semantic drift**: LLM-generated rubrics tend to enforce rigid checklists rather than recognizing semantically equivalent content, thereby penalizing valid alternative answers.

To investigate these limitations empirically, we compared our expert-designed rubrics against two automated baselines: generic dimension-based rubrics (Rubric_{dim}) following Du et al. (2025), and LLM-generated rubrics based on reference materials (Rubric_{llm}). Our experiments demonstrate that Rubric_{dim} exhibits severe ceiling effects where all systems achieve 87-98% coverage regardless of actual quality, while Rubric_{llm} demonstrates low recall (0.60) against human annotations due to excessive rigidity. In contrast, our expert-designed rubrics achieve superior human alignment according to Section 4.3, effectively balancing rigor with flexibility for open-ended research questions. Detailed experimental results and analysis are provided in Appendix I.

Insight Assessment vs. Pairwise Comparison Pairwise comparison is an evaluation method where a judge model directly compares two responses side-by-side to determine which is better (Zheng et al., 2023). It is widely adopted for its ability to leverage models’ discriminative capabilities over generative scoring, and elimination of subjective scaling biases through two-choice comparison (Chiang et al., 2024; Dubois et al., 2025). However, this evaluation method faces a critical limitation of **length bias problem**, especially in open-ended research evaluation (Sharma et al., 2025).

Our analysis reveals that an extremely strong positive correlation is observed between response length and pairwise win-rate ($r \approx 0.95$, $p < 0.05$), indicating a severe length bias. This result raises significant concerns about its reliability for assessing the quality of DARS response. In contrast, our rubric-based assessment mitigates this bias by evaluating the presence of specific, expert-verified criteria, which are independent of response length. Details in pairwise comparison analysis are provided in Appendix G.7.

B Limitations and Future Works

B.1 Limitations

While ResearcherBench represents a significant advance in evaluating DARS capabilities on open-ended AI research tasks, several limitations should be acknowledged.

Domain Specificity. Our benchmark focuses on AI-related research questions, which may limit the generalization of findings to other scientific domains. The specialized nature of AI research questions may exhibit different characteristics compared to research questions in other fields, and DARS performance patterns might vary across disciplines.

Black-box Commercial Systems. The black-box nature of commercial DARS systems prevents deeper analysis of why certain systems excel in specific question types or what design principles enable superior AI research assistance.

B.2 Future Works

Several promising directions emerge from our findings that warrant further investigation.

Cross-Domain Expansion. Expanding ResearcherBench to additional scientific domains would provide valuable insights into the domain-generalizability of DARS capabilities. Developing comparable benchmarks for fields such as biology, chemistry, physics, and social sciences would enable cross-domain analysis of research assistance patterns and reveal whether the open consulting superiority we observed generalizes beyond AI research.

Continuous Benchmark Evolution. As open-ended AI research rapidly evolves, maintaining the relevance and currency of our benchmark requires periodic incorporation of new research questions that reflect the latest developments in AI and related fields. This ongoing evolution will ensure that ResearcherBench remains aligned with the cutting-edge of scientific inquiry, capturing emerging research paradigms and novel challenges that push the boundaries of current knowledge. Regular updates will involve collaboration with active researchers to identify and validate the most pressing and innovative questions in contemporary AI research.

Longitudinal DARS Evaluation. Implementing systematic longitudinal evaluation of DARS systems would provide crucial insights into capability development trajectories and technological advancement patterns. By continuously assessing new DARS releases and iterations on our benchmark, we can track performance changes over time, identify which aspects of AI research assistance improve most rapidly, and analyze the developmental pathways of different system architectures. This longitudinal analysis would inform both research priorities and development strategies, while providing the community with empirical evidence of progress in AI research assistance capabilities.

These future directions would collectively advance our understanding of AI research assistance capabilities and contribute to the development of systems that can serve as research partners in scientific discovery across diverse domains.

C Dataset Construction

C.1 Definition of Question Types

Table 5 shows detailed definition and example of each question type.

C.2 Details in Data Collection

The initial corpus of 932 candidate questions is from three primary sources: (1) **Laboratory research discussions (255 questions)**: We collected these questions from transcripts of laboratory discussions across various research topics, where researchers actively grapple with unsolved technical challenges; (2) **Interviews with leading AI researchers (344 questions)**: We gathered these questions from the subtitles of frontier interviews conducted on video platforms (such as YouTube), which often reveal emerging research directions; and (3) **Scientific forum discussions (333 questions)**: These questions were collected from research-oriented topics on online forums like X (formerly Twitter) and Zhihu, where researchers share their latest work and insights.

Data collection was conducted between October 2024 to April 2025. We also gathered the contextual information of each question, which serves as supporting material for the rubric construction process.

C.3 Details in Question Selection

We formalized the tailored criteria of different question types into structured rubrics using 5-point Likert scales (1=Poor, 5=Excellent) (Joshi et al., 2015). The detailed criteria and

Table 5: Example of Different Question Types. Each question type follows a detailed definition and an example. Technical detail questions emphasize precise verification, literature review questions focus on comprehensive synthesis, and open consulting questions prioritize broader insights.

Question Type	Definition	Example
Technical Details	Questions requiring explanations of methodologies, implementations, or theoretical concepts with a strong emphasis on accuracy and verification.	Why doesn't ChatGPT directly fine-tune using Reward-Model data, but instead use RLHF? Give me a more deep technical report, and focus on references to recent research papers on this topic.
Literature Review	Questions that involve synthesizing findings from multiple research papers, comparing methodologies, and identifying trends or gaps in existing literature.	For complex reasoning tasks, what are the strengths of current agent technologies, and what are their limitations? Please analyze this in the context of research since June 2024.
Open Consulting	Questions that explore emerging trends, strategic insights, detailed solutions or broader implications, often requiring subjective interpretation and expert judgment.	Could transformer architectures be fundamentally reimagined to process multimodal inputs (e.g. video, audio, or text) with the same efficiency they process text?

specific rubrics for each question type are presented as follows. These criteria served as guidelines for human expert annotators in the question selection process from the initial corpus of 932 questions. Our final result shows that inter-annotator agreement (Pearson Correlation) achieved $r = 0.76$, and only questions achieving average scores ≥ 4.5 were retained.

Technical Details Questions Evaluation Rubric

Key Evaluation Dimensions:

- **Technical Specificity:** Whether the question targets specific technical concepts, implementations, or methodologies with clear scope and boundaries.
- **Precision Requirements:** Whether the question demands accurate, detailed explanations that can differentiate between correct and incorrect technical knowledge.
- **Factual Verifiability:** Whether the answer can be verified against authoritative sources (documentation, standards, publications) with objective criteria.
- **Application Value:** Whether the question effectively reflects the model's retrieval capability for fine-grained technical problems.

Scoring Standards:

- **5 points:** Question pinpoints specific technical details with clear scope; requires precise, accurate explanations that demonstrate deep technical understanding; answers are easily verifiable against authoritative sources and remain stable over time; excellently tests model's ability to retrieve and explain complex technical knowledge.
- **4 points:** Question targets specific technical aspects with good clarity; requires detailed explanations with minor ambiguity; answers are generally verifiable with stable technical foundations; effectively tests technical knowledge retrieval with room for minor improvements.

- **3 points:** Question addresses technical details but with moderate specificity; requires explanations that may allow some interpretation; answers are partially verifiable but may depend on context or evolving standards; adequately tests technical knowledge but could be more precise.
- **2 points:** Question lacks technical specificity or is too broad; explanations may be vague or surface-level; answers are difficult to verify objectively or depend heavily on time-sensitive information; limited effectiveness in testing precise technical knowledge.
- **1 point:** Question is technically vague or overly general; fails to demand specific technical knowledge; answers cannot be objectively verified or are highly dependent on rapidly changing information; ineffective for testing detailed technical understanding.

Literature Review Questions Evaluation Rubric

Key Evaluation Dimensions:

- **Research Direction Clarity and Survey Scope:** Whether the question provides clear research direction and can guide comprehensive literature comparison and synthesis across multiple perspectives and methodologies.
- **Literature Coverage Requirements:** Whether the question demands systematic exploration of key papers, major approaches, and important findings in the specified research area.
- **Verifiability and Stability:** Whether the research direction is well-established with literature that is easily retrievable and verifiable, maintaining relevance over time.

Scoring Standards:

- **5 points:** Question provides clear research direction with well-defined scope; guides comprehensive literature exploration across multiple dimensions; targets stable research area with abundant, easily accessible literature; excellent potential for meaningful survey output.
- **4 points:** Question offers generally clear direction with adequate scope definition; encourages broad literature coverage; targets established research area with good literature accessibility; solid foundation for comprehensive survey work.
- **3 points:** Question provides moderate direction clarity with acceptable scope; requires literature coverage but may lack depth requirements; targets reasonably stable research area with moderate literature accessibility; sufficient for basic survey objectives.
- **2 points:** Question direction is somewhat unclear or too narrow/broad; limited guidance for comprehensive literature exploration; targets area with limited or hard-to-access literature; marginal value for survey purposes.
- **1 point:** Question lacks clear research direction or proper scope definition; fails to guide systematic literature exploration; targets unstable area with poor literature accessibility; inadequate for quality survey work.

Open Consulting Questions Evaluation Rubric

Key Evaluation Dimensions:

- **Openness and Depth:** Whether the question encourages creative exploration of multiple perspectives and stimulates deep, multi-dimensional thinking beyond conventional approaches.
- **Forward-Looking Value:** Whether the question addresses emerging trends, future challenges, or strategic insights that provide meaningful guidance for research directions or industry development.
- **Conceptual Innovation Potential:** Whether the question can inspire novel viewpoints, creative problem-solving approaches, or innovative thinking that advance understanding in the field.
- **Balanced Grounding and Long-term Relevance:** Whether the question maintains reasonable connection to existing knowledge while avoiding over-dependence on transient trends, ensuring lasting value.

Scoring Standards:

- **5 points:** Question demonstrates exceptional openness that encourages creative exploration across multiple dimensions; addresses significant forward-looking challenges or strategic opportunities; inspires innovative thinking and novel approaches; maintains strong grounding in fundamental principles while offering lasting relevance beyond current trends.
- **4 points:** Question shows high openness with good potential for multi-perspective exploration; addresses meaningful future-oriented topics with strategic value; encourages innovative thinking with reasonable grounding; demonstrates good long-term relevance.
- **3 points:** Question provides moderate openness with some potential for creative exploration; addresses topics with acceptable forward-looking value; allows for some innovative thinking but may lack depth; maintains reasonable balance between innovation and grounding.
- **2 points:** Question offers limited openness with constrained exploration potential; addresses topics with minimal strategic or future value; provides little inspiration for innovative thinking; may be either too abstract or too tied to current trends.
- **1 point:** Question lacks meaningful openness and fails to encourage creative exploration; addresses topics with little strategic value or future relevance; provides minimal potential for innovative thinking; either completely ungrounded or overly dependent on temporary trends.

C.4 Examples of Question Selection

To illustrate the application of our evaluation framework, we present concrete examples below to demonstrate how questions are assessed across different categories and quality levels. These examples showcase the evaluation process in action, highlighting both high-quality questions suitable for benchmark inclusion and problematic questions that require significant improvement.

Example 1

Question: Compare the Transformer and Mamba model architectures, analyzing their performance and technical characteristics in different application scenarios. Based on the latest research, discuss the advantages and disadvantages of both models and their applicable scenarios.

Question Type: Literature Review

Annotator:

1. Research Direction Clarity and Survey Scope

- **Score:** 4/5
- **Explanation:** The question provides generally clear research direction with a well-defined comparative framework between two specific model architectures (Transformer and Mamba). However, the scope could be more precise regarding specific application domains, evaluation metrics, or time frame for "latest research."

2. Literature Coverage Requirements

- **Score:** 5/5
- **Explanation:** The question effectively demands systematic exploration of foundational papers on both architectures (Attention is All You Need for Transformers, Mamba papers), comparative studies and benchmarks and Recent advances and variants in both families.

3. Verifiability and Stability

- **Score:** 3/5
- **Explanation:** Transformer architecture is extremely well-established with abundant, easily accessible literature, and core architectural principles are stable and well-documented. However, Mamba is a relatively recent architecture (2023), so the literature base is still developing. Rapid evolution means some findings may become outdated quickly.

Overall Assessment:

Overall Score: 4/5

Recommendation: HIGH VALUE for inclusion

Explanation: This question demonstrates high benchmark value by combining real research relevance, comprehensive skill testing across multiple dimensions (synthesis, analysis, reasoning, application), balanced complexity, contemporary significance in architectural debates, and objective verifiability against established research standards.

Example 2

Question: What are the differences between GQA and MQA? And what are their respective applicable scenarios?

Question Type: Technical Details

Annotator:

1. Technical Specificity

- **Score:** 2/5
- **Explanation:** The question lacks sufficient technical specificity. Firstly, "GQA" is ambiguous - could refer to Grouped Query Attention or Graph Question Answering. Secondly, the scope is too broad, asking for general differences and applications, no specific implementation details, parameters, or technical boundaries are defined. Last but not least, context about which domain is missing (NLP, computer vision, etc.).

2. Precision Requirements

- **Score:** 2/5
- **Explanation:** The question has significant precision issues: The ambiguous "GQA" term makes precise explanation difficult, and respective application scenarios is too general. It doesn't demand specific technical details about attention mechanisms, computational complexity, or implementation specifics, and allows for surface-level explanations rather than deep technical understanding.

3. Factual Verifiability

- **Score:** 3/5

- **Explanation:** If GQA refers to Grouped Query Attention, it can be verified against recent research papers, while MQA is well-documented in transformer literature. However, the ambiguous terminology makes verification challenging.

4. Application Value

- **Score:** 2/5
- **Explanation:** Limited application value due to the ambiguity reduces its effectiveness in testing precise technical knowledge. This question doesn't effectively differentiate between surface knowledge and deep understanding, and it's too general to test fine-grained technical retrieval capabilities.

Overall Assessment

Overall Score: 2.25/5

Recommendation: LOW VALUE for inclusion

Explanation: This question should NOT be included due to fundamental ambiguity ("GQA" is unclear), poor discriminative power, verification challenges, and limited technical rigor that fails to effectively test precise technical knowledge.

To improve this question, it should clearly specify "Grouped Query Attention (GQA)" vs "Multi-Query Attention (MQA)", demand specific technical details (computational complexity, memory usage, implementation differences), and define clear evaluation metrics with concrete examples.

D Guidelines for Rubric Design

D.1 Key Insight Extraction

We firstly use Claude-3.7-Sonnet to analyze the contextual source material for each question, identifying key insights and generating comprehensive reference materials. For different question types, we design different focus areas when extracting key insights:

Technical Details Questions - Key Insight Guidelines

For technical details questions, the analysis focuses on extracting structured key insights with emphasis on:

1. Precise technical specifications and parameters
2. Detailed algorithmic descriptions and mathematical formulations
3. Implementation considerations and computational requirements
4. Performance metrics and efficiency analyses
5. Technical limitations and edge cases
6. Optimization techniques and fine-tuning procedures
7. Code examples or pseudocode where applicable
8. System architecture and component interactions
9. Technical dependencies and environmental requirements
10. Debugging approaches and common technical pitfalls

Literature Review Questions - Key Insight Guidelines

For literature review questions, the analysis focuses on extracting structured key insights with emphasis on:

1. Comprehensive overview of the technological landscape

2. Historical development and evolution of relevant technologies
3. Current state-of-the-art approaches and methodologies
4. Comparative analysis of different technical solutions
5. Key research papers, influential publications and bibliographic references
6. Emerging trends and future research directions
7. Major contributors and research groups in the field
8. Theoretical foundations and fundamental principles
9. Cross-disciplinary connections and applications
10. Benchmark datasets and evaluation frameworks commonly used in the field

Open Consulting Questions - Key Insight Guidelines

For open consulting questions, the analysis focuses on extracting structured key insights with emphasis on:

1. Provision of new insights beyond common knowledge or existing literature
2. In-depth analysis of the question from multiple perspectives
3. Critical thinking and identification of key challenges and core problems
4. Novel hypotheses, conceptual frameworks, or alternative viewpoints
5. Strategic discussions on potential research directions or practical solutions
6. Integration of cross-disciplinary knowledge to enrich the analysis
7. Reflection on the broader implications, including societal, ethical, and industrial impacts
8. Exploration of future trends and transformative opportunities
9. Expert judgment supported by logical reasoning and evidence
10. Creative and thought-provoking ideas that inspire further discussion

D.1.1 Key Insight Extraction Prompt

We employ the following prompt to extract key insights from the contextual source material of questions, and generate auxiliary materials as reference.

Key Insight Extraction Prompt Template

<system_role>

You are an expert research analyst specializing in extracting high-value insights from academic and technical content.

</system_role>

<user_prompt>

Your task is to identify and structure key insights that demonstrate deep understanding and expert-level analysis. Given the following question and its contextual source material, extract key insights following the specific guidelines for {Question Type} questions.

Question: {Question}

Source Material: {Source Context}

Guidelines: {Guideline for Question Type}

Please identify and extract 8-15 key insights that represent the most valuable and insightful aspects of addressing this question. Each insight should be:

- Substantive and demonstrate deep understanding
- Directly relevant to answering the question
- Represent expert-level analysis or specialized knowledge
- Be specific enough to be evaluable

Format your response as a structured list of key insights with brief explanations.

</user_prompt>

Reference Material Generation Prompt Template

<system_role>

You are an expert research analyst specializing in extracting relevant information from the document and providing a comprehensive reference materials.

</system_role>

<user_prompt>

Your task is to:

- Carefully analyze the document script provided
- Extract all relevant information related to the specific question, containing all the key insights
- Organize the information in a coherent, well-structured response
- Provide accurate, helpful information based primarily on the document content

IMPORTANT: If the document and those key insights do not contain enough information to fully answer the question, you may supplement with your general knowledge, but you must clearly indicate which parts are from the document and which parts are your additional context or expertise. Always prioritize information from the document and provided key insight, and only add relevant knowledge when necessary to provide a more complete answer.

Question: {Question}

Full Document: {Document}

Key Insights: {Key Insights List}

Please provide a comprehensive answer to the question based primarily on the information in the document and key insights. If needed, you may supplement the answer with your own knowledge, but clearly distinguish between information from the document and your additional insights.

</user_prompt>

D.1.2 Human Verification and Rubric Design

Based on the extracted key insights and reference materials, we invite expert annotators to design corresponding rubrics for each question. We established the following annotation guidelines:

Rubric Design Annotation Guidelines

Task Overview:

You will be presented with a research question, a list of reference materials, and complete contextual information discussing this research question. Your task is to design rubric based on reference materials for evaluating the quality of reports answering this research question, and assign weights to different criteria based on value judgment.

Rubric Design Requirements:

- **Clarity and Verifiability:** Each criterion must be clearly described and easy to verify objectively.
- **Independence:** Ensure each criterion is independently verifiable without overlap.
- **Conceptual Focus:** Focus on essential concepts rather than specific examples.
- **Evaluative Phrasing:** Phrase each criterion as an evaluative statement, not a descriptive one.
- **Objective Assessment:** Make each criterion specific enough to enable clear pass/fail evaluation.
- **Action-Oriented Language:** Use verbs like "Explains," "Describes," "Discusses," "Outlines," "Provides," "Analyzes," "Compares," "Identifies" to indicate the expected level of detail and engagement with concepts.
- **Contextual Reference:** The reference materials serve only as auxiliary annotation text and may not accurately reflect the original discussion content. When confused, refer to the original context for specific information.

Weight Assignment Guidelines: Assign weights from 1-3 based on the importance of each criterion to answering the question comprehensively:

- Higher weights (3) should be assigned to criteria that are core to understanding the core question
- Medium weights (2) for supporting criteria that add depth and context
- Lower weights (1) for nice-to-have criteria that enhance but are not essential to the answer
- Ensure total weights reflect the relative importance hierarchy of different aspects

Quality Control Measures:

- Each criterion should be binary assessable (present/absent)
- Avoid subjective language that could lead to inconsistent evaluation
- Test each criterion against the reference materials to ensure it captures meaningful distinctions
- Ensure each criterion collectively cover the essential aspects of a comprehensive answer

D.2 Rubric Example

To demonstrate the dynamic rubric generation process, we present examples showing how specific evaluation criterion are designed based on key insights extracted from responses.

The question is: How should the development of generative AI evolve: focusing on dialogue-based systems (Chat) or autonomous action-taking systems (Agent)? What are the key differences, technological requirements, and future implications of each approach?

Example 1**Key Insight:**

Agent evolution mirrors historical tech patterns like smart speakers becoming home automation hubs and Robotic Process Automation (RPA) transitioning to Agentic Process Automation (APA).

Criterion:

Provides relevant historical analogies to similar technological evolutions (e.g., smart speakers, voice assistants). (weight=1)

Example 2**Key Insight:**

Chat systems focus on information processing and linguistic exchange, operating as ‘brain and mouth’ entities confined to conversational domains (e.g. ChatGPT), while agent systems represent a paradigm shift from predefined logic in traditional software to autonomous, goal-oriented solutions generated collaboratively between developers and AI.

Criterion:

Clearly distinguishes between process-oriented (Chat) and goal-oriented (Agent) frameworks. (weight=2)

The first criterion captures the importance of historical context and analogical reasoning, while the second focuses on the fundamental architectural distinctions between different AI paradigms.

E Rubric Statistics

To show rubric diversity, we provide a comprehensive quantitative overview of the criteria count and weight distributions across all 65 benchmark questions.

E.1 Criteria Count Distribution

Figure 4 shows the distribution of the number of criteria per question. The majority of questions (81.5%) contain between 11 and 18 criteria, reflecting the complexity of open-ended AI research questions.

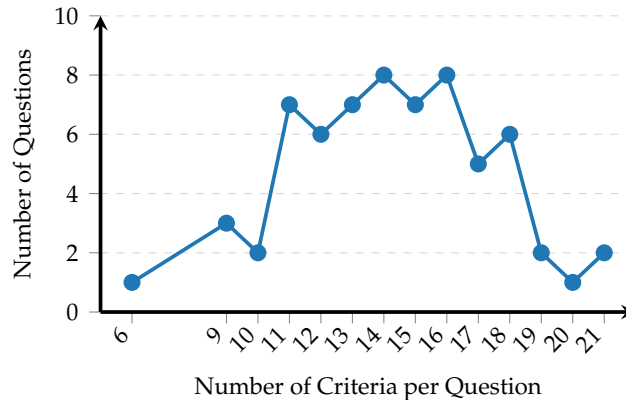


Figure 4: Distribution of criteria count per question (65 questions total). Most questions have 11–18 criteria (mean = 14.3).

E.2 Weight and Per-Category Breakdown

Each criterion is assigned an importance weight $w_i \in \{1, 2, 3\}$ by domain experts: weight 1 for nice-to-have insights, weight 2 for supporting criteria, and weight 3 for core insights essential to answering the question. Across all 931 criteria, weight-2 criteria dominate (50.9%), consistent with our rubric design philosophy that most criteria should reflect important but not indispensable insights. Weight-3 criteria (13.6%) are reserved for the most central concepts that any high-quality answer must address.

Table 6 disaggregates these statistics by question type. Technical Details questions have slightly more criteria on average and a higher proportion of weight-3 (core) criteria, reflecting their requirement for precise, verifiable knowledge.

Table 6: Rubric statistics broken down by question category, including criteria count, total score range, and weight distribution (w=1: nice-to-have; w=2: supporting; w=3: core).

Category	#Q	Criteria Count		Total Score		Weight Dist.		
		Range	Mean	Range	Mean	w=1	w=2	w=3
Literature Review	20	6–21	14.8	14–39	25.6	40%	48%	13%
Open Consulting	33	9–21	13.8	15–39	24.7	34%	54%	13%
Technical Details	12	11–18	15.0	19–38	27.7	33%	49%	18%
Overall	65	6–21	14.3	14–39	25.5	35%	51%	14%

F Experimental Setup

Report Generation Prompt Template for Non-DARS systems

Generate a comprehensive research report addressing the following research questions. Your report must include the clear structure, detailed explanations, and references to relevant academic sources. You should use IEEE style citations for all references, Use numbered citations in square brackets like [1], [2], [3].
Research Question: [QUESTION]

For most DARS systems except Perplexity Deep Research, interaction was only possible through web user interfaces (WebUI), which somewhat limited the scalability of our evaluation experiments. To ensure fairness across different DARS systems, we employed default settings when interacting with all systems. Specifically, OpenAI Deep Research, Claude Research and Doubao Deep Research typically asks follow-up questions after the initial user query, and our standard response was "By default." to maintain consistency across evaluations. Gemini-2.5-Pro Deep Research pre-generates research plans before conducting research, and we directly used these generated plans without any modifications to maintain system autonomy and avoid human intervention bias.

For non-DARS systems with web search capabilities (GPT-4o Search Preview and Perplexity: Sonar Reasoning Pro), we configured the Search Context Size to "High" setting to maximize their research capabilities and ensure fair comparison with dedicated research agents. This enhancement was designed to compensate for the lack of specialized research workflows in these systems by providing them with expanded search context, thereby enabling more comprehensive information retrieval and analysis. Additionally, we designed a unified prompt for non-DARS systems to guide models in generating responses with proper report format.

G Experimental Details

G.1 Data Collection Timeframes

Due to the lack of transparency regarding model iterations and technical details in most commercial DARS systems, we explicitly documented the timeframes during which our data collection is taken in Table 7. We document evaluation windows transparently so that the community can accurately interpret results in the context of each system’s version at evaluation time, and re-evaluate any system using our open-source framework as needed.

DARS	Data Collection Timeframe
OpenAI Deep Research	March 24, 2025 – April 29, 2025
Perplexity Deep Research	March 24, 2025 – April 15, 2025
Grok3 Deep Search	March 25, 2025 – April 14, 2025
Grok3 Deeper Search	April 18, 2025 – April 19, 2025
Gemini Deep Research	April 15, 2025 – April 21, 2025
Claude Research	July 21, 2025 – July 28, 2025
Doubao Deep Research	July 21, 2025 – July 28, 2025
Mita Deep Research	July 21, 2025 – July 30, 2025

Table 7: Data Collection Timeframes for DARS

G.2 Implementation Details

While our evaluation framework follows the methodology outlined in Section 4, several specific optimizations were adopted to enhance robustness and efficiency in practice.

Insight Assessment Implementation. For insight assessment, we evaluate each criterion independently using the prompt template provided in Appendix G.3.

Citation Assessment Implementation. For citation assessment, we implemented a section-based claim extraction strategy rather than processing entire reports simultaneously, considering computational efficiency and context length limitations. Specifically, reports are segmented into sections by paragraph boundaries, and claims are extracted from each section independently using the template in Appendix G.4. Finally, extracted claims from all sections are subsequently aggregated to form the complete claim set for each response.

Following extraction, claims sharing identical URL sources are grouped to optimize verification and reduce redundant web retrieval. Each URL group undergoes batch verification using the prompt in Appendix G.5. To handle potential failures and errors robustly, we extended the binary judgment framework into a three-way classification (“yes”, “no”, and “unknown”). This three-way classification allows the system to gracefully handle cases where claims are incompletely extracted or source URLs are inaccessible. Claims classified as “unknown” are excluded from final metrics to ensure evaluation reliability.

G.3 Rubric Coverage Evaluation Prompt

Rubric Coverage Evaluation Prompt

Task

Determine whether the AI response adequately covers the specific criterion provided. Answer with “yes” or “no” followed by a brief justification.

Input Materials

<Question>: {question}

<Criterion>: {criterion}

<Weight>: {weight} (indicates the importance of this criterion)

<AI Response>: {ai_response}

Evaluation Criteria

- Answer “yes” if the AI response clearly includes or adequately expresses the main content of the criterion
- Answer “yes” if the response conveys the same meaning as the criterion, even if using different terminology or phrasing
- Answer “no” if the AI response only partially addresses or completely fails to mention the content of the criterion
- Consider semantic equivalence, not just keyword matching
- Pay special attention to technical details, numerical values, and specific claims in the criterion

Output Format

Your answer must begin with either “yes” or “no” followed by a brief justification.

Example format

“yes: The response clearly addresses this criterion by explaining [specific detail]...”

“no: While the response mentions [related concept], it fails to address [specific aspect] of the criterion...”

G.4 Claims Extraction Prompt

Claims Extraction Prompt

Task Description

Extract all factual claims from the provided academic paper. Each claim should be a factual statement that can be verified. Claims may or may not have supporting citations.

Input

A Research Question and a complete academic paper containing factual claims, some of which may have citation markers and corresponding URLs (inline or in reference section).

Output Requirements

- Extract each distinct factual claim throughout the entire paper
- For each claim, output a JSON object with:
 - The exact claim text as a string
 - The original text from the paper containing this claim (context)
 - The corresponding citation URL as source (if a citation marker directly follows the claim)
- If a claim has a citation marker directly following it, return the supporting URL
- If a claim have no citation marker directly following it, return an empty string
- Ensure all string values are properly escaped for valid JSON format (e.g. Replace internal quotation marks (") with escaped quotation marks (\"))
- Return a JSON array containing all claim objects

Format Specification

```
[
  {
    "claim": "The exact statement representing a factual claim",
    "context": "The original sentence or passage from the paper
                containing this claim",
```

```

    "source": "https://example.com/source1"
  },
  {
    "claim": "Another factual statement without direct citation",
    "context": "The original sentence or passage from the paper
                containing this claim",
    "source": ""
  }
]

```

Guidelines for Claim Identification

1. A claim should be a complete, standalone factual statement
2. Maintain the original wording where possible, but remove unnecessary context
3. Extract all factual claims regardless of whether they have citation support
4. Only consider to map citation markers (numbers, author names, etc.) to their corresponding URLs in the references section when it directly follow the claim statement
5. Exclude opinions, speculations, or methodological descriptions
6. Extract the context passage containing each claim for verification purposes
7. If multiple claims are associated with the same citation, extract them as separate entries

Citation URL Mapping

- If URLs appear directly after claims, use those URLs directly
- Citation markers (e.g. follows a number or [number]) must directly follow the claim to be considered as supporting that claim
- If claims use citation markers that reference a bibliography or reference section, locate the corresponding URLs in that section
- If a claim has no directly following citation marker, use an empty string for source

Please extract all claims from the response content and output in the specified JSON format:

Research Question: [QUESTION]

Response Content: [CONTENT]

References: [REFERENCES]

G.5 Claim Verification Prompt

Claim Verification Prompt

Task Description

Your task is to verify whether claims are supported by the provided reference content.

Input

- A reference content that contains supporting information
- A list of claim-context pairs that need to be verified against the reference

Output

For each claim, respond with 'yes', 'no', or 'unknown' to indicate whether the claim is supported by the reference content. Output in the specified JSON format.

Output Format Specification

```

[
  {

```

```

    "id": 1,
    "result": "yes"
  },
  {
    "id": 2,
    "result": "no"
  },
  {
    "id": 3,
    "result": "unknown"
  }
]

```

Verification Guidelines

Claim Support Determination

If the reference is valid, for each given claim:

- **‘yes’**: If the facts or data in the claim can be found entirely or partially within the reference content
- **‘no’**: If all facts and data in the statement cannot be found in the reference content
- **‘unknown’**: If verification encounters difficulties (such as semantic incompleteness, ambiguity, or other issues that make verification impossible), or reference contents are not available (‘page not found’ message, connection errors, or other non-content responses)

Claims must be verifiable from the content provided, not based on general knowledge.

Using Context Information

If you encounter difficulties when verifying claims (e.g., semantic incompleteness/ambiguity issues), refer to the corresponding additional context. If problems still exist after considering the paragraph context, output ‘unknown’.

Please provide your verification results in the specified JSON format.

Source: [SOURCE]

Claim-Context Pair List: [CLAIM_LIST]

G.6 Error Analysis and Examples

G.6.1 Taxonomy and Methodology

To gain deeper insights into the failure modes of DARS, we perform a systematic error analysis on a representative subset of our dataset. We categorize common failure patterns along two dimensions: insight assessment errors and citation assessment errors:

Insight Assessment Errors. For responses that fail to satisfy any evaluation criteria, we identify two primary error categories:

- **Incompleteness**: The response completely fails to address or satisfy the criterion, demonstrating a fundamental gap in research coverage.
- **Partial Missing**: The response addresses only a subset of the criterion’s requirements, indicating incomplete analysis or synthesis.

Citation Assessment Errors. For claims in the responses that fail verification, we also identify two primary error categories:

- **Invalid Attribution**: The cited URL is expired, malformed, fabricated, or otherwise inaccessible, preventing verification regardless of claim validity.
- **Unsupported Answer**: The URL contains relevant information but the model misrepresents, misinterprets, or hallucinates content not present in the source.

Error Type	Open Consulting	Literature Review	Technical Details
<i>Insight Assessment</i>			
Incompleteness	4.47	6.23	5.63
Partial Missing	1.80	2.64	2.93
<i>Subtotal</i>	<i>6.27</i>	<i>8.87</i>	<i>8.56</i>
<i>Citation Assessment</i>			
Invalid Attribution	2.48	2.35	1.42
Unsupported Answer	16.40	16.35	15.50
<i>Subtotal</i>	<i>18.88</i>	<i>18.70</i>	<i>16.92</i>

Table 8: Average number of errors per response across different question types. Error distributions reveal distinct failure patterns in rubric coverage versus factual grounding.

Human annotators examined responses from five representative DARS (OpenAI Deep Research, Gemini Deep Research, Claude Research, Doubao Deep Research, and Mita Deep Research), with 10 responses sampled from each system (50 responses in total). For insight assessment, annotators directly categorized uncovered criteria according to the three-way taxonomy. For citation assessment, annotators first employed the same web retrieval API used in our main experiments to determine URL accessibility (Invalid Attribution), then compared model responses against retrieved content to identify hallucinations (Unsupported Answer).

G.6.2 Results and Analysis.

Table 8 presents the average error count per response across question types. Several key patterns emerge from this analysis:

Error Check for Insight Assessment. 1) Open consulting questions exhibit the lowest average error rate (6.27), corroborating our main findings in Section 5.2 regarding DARS’s outstanding performance on exploratory research tasks. 2) Literature review questions show the highest rubric error rate (8.87), particularly in incompleteness (6.23), indicating that systematic coverage of research breadth remains challenging even for state-of-the-art DARS. 3) Incompleteness errors dominate partial missing errors across all question types (2.5:1 of rubric errors on average), suggesting that DARS demonstrate strong integration and reasoning capabilities for information they successfully retrieve, but struggle to answer when relevant information is not retrieved during the search phase. This indicates that retrieval capability might represent a primary bottleneck in insight assessment.

Error Check for Citation Assessment. 1) DARS demonstrate higher hallucination rates in open consulting (16.40) and literature review questions (16.35) compared to technical details questions (15.50). This pattern suggests that when synthesizing broad conceptual insights or surveying literature landscapes, DARS are more prone to misrepresenting source content than when retrieving specific technical information. 2) Failure modes in the citation assessment varies significantly across different systems. For instance, Mita Deep Research exhibits a strong bias due to inaccessible URLs (Invalid Attribution : Unsupported Answer = 2.32 : 15.56), suggesting retrieval quality issues in its web search engine. Conversely, Doubao Deep Research demonstrates a different failure profile (Invalid Attribution : Unsupported Answer = 1.98 : 16.36), with hallucination problems primarily stemming from the model’s content generation rather than source accessibility.

G.6.3 Example Cases of Error Types

To illustrate the error patterns identified in our analysis, we present representative examples from our annotated dataset. These cases demonstrate the specific manifestations of each error category and provide actionable insights into future DARS system improvements.

Incompleteness Error Example: Claude Research

Question: What new types of ‘creative infrastructure’ does the web need to support AI-generated 3D/immersive experiences while maintaining open standards?

Criterion: Discuss how the transition from Developer Experience (DX) to Agent Experience (AX) represents a fundamental shift in web development philosophy for supporting autonomous content generation.

Error Analysis: The response never engaged with the specific conceptual shift from Developer Experience (DX) to Agent Experience (AX). Instead, it focused solely on technical and infrastructural challenges of AI-generated 3D web content, ignoring the rubric’s requirement to discuss how AX fundamentally redefines web development philosophy for autonomous content generation.

Partial Missing Error Example: Gemini Deep Research

Question: In multimodal pretraining, the current mainstream paradigms are based on image tokens and stable diffusion. Analyzing the latest advancements (by April 2025) in these two technical approaches, with reference to the most recent papers, which one appears to be more promising and why?

Criterion: Explains how compressed latent space operation balances computational efficiency with detail preservation in recent stable diffusion approaches.

Error Analysis: The model response gestured at the idea of operating in a compressed latent space for efficiency but stopped short of explicitly addressing the core trade-off that the rubric required: how recent stable diffusion methods balance computational gains with the preservation of fine-grained detail. It mentioned latent diffusion and compression, but failed to connect this to concrete aspects of detail preservation.

Invalid Attribution Error Example: Doubao Deep Research

Question: “Why can models trained on synthetic data outperform the models that provide the synthetic data? Please find the latest research papers...”

Claim: “Structured Domain Randomization outperformed real data on the KITTI object detection task in Computer Vision.”

Cited Source: https://m.zhangqiaokeyan.com/academic-conference-foreign-international-conference-robotics-and_thesis/020512349961.html

Source Status: Invalid URL (403 Forbidden)

Error Analysis: The cited URL returns a 403 Forbidden error and does not provide access to the source content. Therefore, it is impossible to verify or refute the claim regarding Structured Domain Randomization’s performance on the KITTI object detection task.

Unsupported Answer Error Example: Mita Deep Research

Question: What is the fundamental reason behind the low cost of DeepSeek V3? Is it due to leveraging data distillation from other ‘teacher models’ (such as OpenAI, Gemini, etc.), or adjustments in training and inference precision algorithms?

Claim: “DeepSeek-V3’s inference pricing is at 1% of GPT-4o’s rates.”

Cited Source: <https://arxiv.org/pdf/2412.19437>

Source Title: DeepSeek-V3 Technical Report

Source Status: Accessible

Error Analysis: The claim states that DeepSeek-V3’s inference pricing is at 1% of GPT-4o’s rates. The retrieved source provides extensive details on DeepSeek-V3’s architecture, training costs, and performance, but does not mention or quantify inference pricing relative to GPT-4o. There is no explicit comparison or percentage given for inference costs between DeepSeek-V3 and GPT-4o in the document.

G.7 Pairwise Comparison Analysis

For experimental rigor, we conducted pairwise comparison to complement our rubric-based assessment. We designed the evaluation prompt following the pairwise comparison methodology from the LLM-as-a-Judge variations proposed in (Zheng et al., 2023). Our pairwise comparison approach involves systematically comparing responses from different DARS models to the same questions, using a judge model to determine which response is better.

G.7.1 Methodology

We conducted pairwise comparisons across all evaluated DARS systems. For each pair of models (Model A and Model B), we prepared their responses to all 65 questions for comparison. The judge model then evaluates each pair of responses side-by-side to determine the winner.

To mitigate potential positional bias that might arise from the presentation order of responses, we implemented a dual evaluation strategy: for each question, we evaluated both "Model A vs Model B" and "Model B vs Model A" comparisons. The final verdict for each question is determined by aggregating results from both evaluation directions, ensuring that presentation order does not influence the judgment outcome.

G.7.2 Prompt Design

The evaluation prompt is designed to assess responses based on deep research characteristics, prioritizing accuracy, depth, comprehensiveness, and technical rigor. The specific prompt template is provided below:

Prompt Template for pairwise comparisons

Please act as an impartial judge and evaluate the quality of the responses provided by two AI research assistants to a deep research question displayed below. This is a Deep AI Research System (DARS) evaluation where you should assess which assistant provides a better research-quality response.

Your evaluation should prioritize the following deep research characteristics:

- **Accuracy and Factual Correctness:** The response should be factually accurate, well-cited (if applicable), and free from hallucinations or misinformation
- **Depth and Insight:** The response should demonstrate deep understanding, critical thinking, and meaningful insights rather than surface-level information
- **Comprehensiveness:** The response should cover all necessary aspects, address complexities, and consider multiple perspectives or approaches
- **Cross-domain Synthesis:** The response should effectively synthesize information across different domains, methodologies, or research areas when relevant
- **Frontier Research Awareness:** The response should demonstrate awareness of cutting-edge research, emerging directions, and open problems in the field
- **Technical Rigor:** For technical questions, the response should show appropriate technical depth, precision, and understanding of underlying principles
- **Practical Relevance:** The response should connect theoretical understanding with practical implications and real-world applications when appropriate

Begin your evaluation by comparing the two responses across these dimensions and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. **Do not allow**

Model	Pairwise Comparison			Win Rate	Avg. Length (chars)
	Win	Tie	Lose		
Gemini Deep Research	527	43	15	90.09%	78, 857
Doubao Deep Research	431	85	69	73.68%	58, 453
OpenAI Deep Research	401	84	100	68.55%	49, 335
Mita Deep Research	317	125	143	54.19%	53, 189
Claude Research	245	109	231	41.88%	15, 183
Perplexity Deep Research	195	103	287	33.33%	13, 113
Grok3 DeeperSearch	156	107	322	26.67%	14, 132
Grok3 DeepSearch	141	121	323	24.10%	14, 380
Perplexity: Sonar Reasoning Pro	40	52	493	6.84%	4, 850
GPT-4o Search Preview	34	47	504	5.81%	6, 718

Table 9: Pairwise comparison results showing detailed statistics for each system across all pairwise comparisons. Top performance is highlighted in bold.

the length of the responses to influence your evaluation - prioritize quality and depth over quantity. Do not favor certain names of the assistants. Be as objective as possible.

After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[Research Question] {question}

[The Start of Assistant A’s Response] {response_a} [The End of Assistant A’s Response]

[The Start of Assistant B’s Response] {response_b} [The End of Assistant B’s Response]

G.7.3 Experimental Results

We evaluated all pairwise combinations across the 10 evaluated systems (8 DARS and 2 LLM systems with web search tools), resulting in 45 model pairs. Each pair comparison was conducted on all 65 questions. Table 9 presents detailed statistics for each model across all pairwise comparisons.

G.7.4 Comparison with Insight Assessment Results

Comparing the pairwise comparison results with our insight assessment results from Section 5.2 (Table 4), we observe several notable discrepancies:

1. **OpenAI and Claude Performance Decline:** In pairwise comparison, OpenAI Deep Research achieves a win rate of 68.55%, ranking third, compared to its top-ranked coverage score of 70.32% in insight assessment. Similarly, Claude Research shows a more significant decline, with a win rate of 41.88% in pairwise comparison, ranking fifth, versus a coverage score of 61.13% (ranking third) in insight assessment.
2. **Doubao Performance Increase:** Doubao Deep Research demonstrates a substantial improvement in pairwise comparison, achieving a win rate of 73.68% and ranking second, compared to its coverage score of 57.54% (ranking fifth) in insight assessment.
3. **Gemini Maintains Top Position:** Gemini Deep Research consistently performs best in both evaluation methods, achieving the highest win rate (90.09%) in pairwise comparison and the second-highest coverage score (69.29%) in insight assessment.

G.7.5 Length Bias Analysis

To investigate the cause of these discrepancies, we analyzed the correlation between response length and pairwise comparison win rates. Table 9 and statistical analysis reveal a **strong positive correlation** between average response length and win rate. Statistical analysis demonstrates that response length and win rate exhibit an extremely strong positive correlation:

- **Pearson correlation coefficient:** $r = 0.9487$ ($p < 0.001$, highly significant)
- **Spearman rank correlation coefficient:** $\rho = 0.9273$ ($p < 0.001$, highly significant)
- **Coefficient of determination:** $R^2 = 0.90$, indicating that response length explains **90% of the variance** in win rates

Despite explicit instructions in the prompt to “**Do not allow the length of the responses to influence your evaluation - prioritize quality and depth over quantity**”, the judge model exhibits a strong bias favoring longer responses. This bias likely emerges indirectly through the evaluation criteria: when assessing dimensions such as “Comprehensiveness” and “Depth and Insight”, longer responses may appear more comprehensive or detailed simply due to their greater volume, even if the additional content lacks substantive value.

G.7.6 Discussion and Implications

The observed length bias in pairwise comparison raises significant concerns about its reliability for assessing DARS quality. Our statistical analysis reveals an extremely high correlation between response length and win rates ($r = 0.9487$, $R^2 = 0.90$), indicating that pairwise comparison results are primarily driven by response length rather than actual research quality.

In contrast, our rubric-based assessment methodology effectively mitigates this length bias problem. The rubric evaluation framework assesses responses based on whether they address specific, expert-designed criteria. Each criterion is evaluated independently using binary judgment (satisfied/not satisfied), and only content that directly addresses the criterion contributes to the score. Under this framework, **longer responses do not automatically receive higher scores**; responses must specifically address the expert-designed criteria to earn points, regardless of their overall length. Even if a response is extremely long, it will not score well unless it effectively addresses the rubric criteria, as irrelevant or redundant content does not contribute to the coverage score.

This fundamental difference explains the discrepancies between pairwise comparison and insight assessment results. DARS like Doubao, which generate longer responses, benefit significantly in pairwise comparison due to the length bias, while DARS like OpenAI and Claude, which provide more concise but high-quality responses, are penalized. The rubric-based approach, by focusing on criterion-specific content rather than overall comprehensiveness, provides a more reliable and objective assessment of actual research insight quality.

To provide a causal robustness check, we further conduct a controlled length-expansion experiment in Appendix L, where Claude Research responses are iteratively expanded without adding new information; rubric coverage does not increase, confirming that our framework is resilient to simple length-inflation attacks.

H Case Study

We conduct a case study to demonstrate how different operational modes produce distinct value in research outputs. We identify three primary response modes: 1) **Information Retrieval Mode**: directly utilizing retrieved information to answer rubric criteria, reflecting high-quality source filtering and information retrieval capabilities; 2) **Internal Knowledge Mode**: leveraging the model’s internal knowledge base to answer rubric criteria, requiring strong model capabilities; and 3) **Reasoning-based Synthesis Mode**: performing inference

based on retrieved information to answer rubric criteria, demonstrating DARS’s information integration and analytical capabilities.

We select the third criterion of Question 51 as example to analyze different DARS response modes:

Example of Question and Evaluation Criterion

Question: “What is the conceptual difference between Mixture of Experts (MoE) in Large Language Models versus traditional recommendation systems, and why do LLMs process tokens rather than entire sentences through individual experts?”

Criterion: “Explains at least one computational efficiency benefit of token-level processing versus sentence-level processing in LLMs.”

The following presents response fragments from different DARS that correctly address this criterion, categorized by their operational modes:

Information Retrieval Mode: Perplexity Deep Research

Response: “**Training Speed:** MoE LLMs with token-level routing have faster training and better or comparable performance than dense LLMs on many benchmarks. (Kirkosyan, 2025)”

Analysis: The response directly quotes retrieved information without substantial reasoning. It precisely locates and extracts key information from external sources, specifically citing the original text from the Neptune.ai blog on Mixture of Experts LLMs.

Internal Knowledge Mode: GPT-4o Search Preview

Response: “**Computational Efficiency:** By focusing on tokens, LLMs can apply parallel processing techniques, enhancing computational efficiency and scalability.”

Analysis: The explanation leverages fundamental knowledge about parallel processing in Transformer architectures. This content is not cited and ‘parallel processing’ is not mentioned in the retrieved sources, indicating the model draws from its internal knowledge base rather than real-time web retrieval.

Single-source Reasoning-based Synthesis Mode: Grok3 DeeperSearch

Response: “**Computational Efficiency and Load Balancing:** Token-level assignment facilitates better load balancing across experts. Since experts can process multiple tokens from different sentences in parallel, it ensures efficient utilization of computational resources. If entire sentences were assigned to experts, longer or more complex sentences could overload specific experts, leading to imbalance. For instance, in Switch Transformers, the top-2 routing ensures that experts are utilized evenly, enhancing efficiency. (Fedus et al., 2022)”

Analysis: The model constructs its explanation using retrieved information as supporting evidence. It introduces “load balancing” as a key advantage and references the Switch Transformers paper to support this inference, demonstrating reasoning over retrieved content.

Multi-source Reasoning-based Synthesis Mode: Gemini Deep Research

Response: “**Computational Efficiency via Conditional Computation:** The core promise of MoE is reducing the computational cost (FLOPs) compared to a dense model with the same total parameter count. (Cai et al., 2025) Token-level routing enables this by activating only a small subset of experts (e.g., 1 or 2 out of 8 or more) for each token. (Wolfe, 2024) This sparse activation drastically cuts down the computation required per token during both training and inference. (Sansevierio et al., 2023)”

Analysis: The model synthesizes information from three distinct sources: drawing the core MoE concept from an arXiv survey, explaining the mechanism via Stack Overflow, and reinforcing outcomes through Hugging Face documentation. This demonstrates comprehensive cross-source knowledge integration.

I Human-Designed v.s. LLM-Generated Rubrics: A Comparative Analysis

As discussed in Section 4, traditional LLM-as-a-Judge evaluation with single overall ratings provides insufficient granularity for assessing complex research responses. Open-ended AI research evaluation requires nuanced assessment across multiple dimensions that single scores cannot capture. To address this limitation, ResearcherBench employs a human-in-the-loop three-stage method, engaging domain experts to construct fine-grained rubrics with specific criteria and assigned weights.

However, this expert-driven approach incurs higher annotation costs compared to fully automated methods. A critical question therefore arises: **can the current LLMs reliably automate this rubric construction process?** While we posit that automated systems struggle to identify valid evaluation dimensions for open-ended AI research tasks, it is essential to provide empirical justification for this claim. In this section, we aim to demonstrate that the expert-designed approach yields tangible performance improvements — specifically regarding alignment with human evaluation — thereby justifying the necessity of human expertise over lower-cost automated baselines.

I.1 Experimental Design

We designed a comparative experiment that evaluates our expert-designed rubrics against two distinct types of LLM-generated baselines. These baselines represent the most common strategies for automated evaluation:

1. **Rubric_{dim}**: Rubrics generated based on four predefined dimensions (Comprehensiveness, Insight, Instruction-Following, and Readability), following the approach used in DeepResearch Bench (Du et al., 2025). This baseline represents a common approach where LLMs generate evaluation criteria based on general quality dimensions.
2. **Rubric_{llm}**: Rubrics constructed following the three-stage method detailed in Section 4, but substituting human experts with Claude-3.7-Sonnet in the Criteria Design (Stage 2) and Quality Control (Stage 3) phases. Specifically, we provided the LLM with identical reference materials, and adapted the human annotation guidelines into prompts to generate evaluation criteria and weights (Stage 2), followed by a self-review process to refine the rubric quality (Stage 3). This baseline tests whether LLMs can effectively replicate the human expert’s rubric design process when given the same reference materials.

For both baseline methods, we used the same judge model (o3-mini) and evaluation pipeline as our insight assessment experiment to ensure fair comparison. We evaluated all DARS responses using these alternative rubrics and compared the results against our expert-designed rubric evaluation.

Model	Rubric _{dim}	Rubric _{llm}	Expert-Designed
<i>Deep AI Research Systems</i>			
OpenAI Deep Research	96.93%, [-1.0%, +1.0%]	65.27%, [-5.2%, +5.0%]	70.32%, [-5.0%, +4.9%]
Gemini Deep Research	97.92%, [-1.7%, +1.2%]	67.06%, [-5.5%, +5.4%]	69.29%, [-5.3%, +5.2%]
Claude Research	96.35%, [-1.1%, +1.0%]	56.16%, [-5.2%, +5.2%]	61.13%, [-5.3%, +5.3%]
Mita Deep Research	97.06%, [-0.8%, +0.8%]	56.28%, [-5.6%, +5.5%]	58.35%, [-5.7%, +5.9%]
Doubao Deep Research	95.64%, [-1.4%, +1.2%]	51.30%, [-5.5%, +5.4%]	57.54%, [-5.7%, +5.5%]
Perplexity Deep Research	95.52%, [-1.2%, +1.1%]	39.21%, [-5.0%, +5.0%]	48.00%, [-5.6%, +5.5%]
Grok3 DeepSearch	93.90%, [-2.2%, +1.9%]	40.74%, [-5.1%, +5.2%]	44.14%, [-5.4%, +5.7%]
Grok3 DeeperSearch	95.26%, [-1.4%, +1.3%]	40.21%, [-5.2%, +5.2%]	43.98%, [-5.6%, +5.6%]
<i>LLM with Search Tools</i>			
Perplexity: Sonar Reasoning Pro	93.58%, [-1.9%, +1.7%]	41.11%, [-5.2%, +5.3%]	46.63%, [-5.6%, +5.6%]
GPT-4o Search Preview	87.34%, [-3.6%, +2.8%]	28.13%, [-4.4%, +4.5%]	35.76%, [-5.1%, +5.2%]

Table 10: Comparison of coverage scores across Rubric_{dim}, Rubric_{llm}, and Expert-Designed Rubrics. We also report 95% bootstrap confidence intervals using 10,000 bootstrap samples.

Rubric Type	Acc.	Prec.	Rec.	F1
Rubric _{dim}	0.91	0.91	0.99	0.95
Rubric _{llm}	0.66	0.83	0.60	0.70
Expert-Designed	0.76	0.85	0.76	0.80

Table 11: Human evaluation alignment metrics for different rubric types.

I.2 Coverage Score Results and Analysis

Table 10 presents a comparative analysis of coverage scores across three distinct evaluation protocols: Expert-Designed Rubrics, Rubric_{dim}, and Rubric_{llm}. The quantitative results highlight substantial disparities in the discriminative power of each method:

Rubric_{dim} exhibits severe ceiling effects. All systems achieve extremely high scores (ranging from 87% to 98%), with a minimal performance gap of only 10.58% between the highest and lowest performing models. This lack of variance indicates that criteria generated purely from generic metrics are overly lenient. They fail to capture nuanced quality differences, effectively categorizing almost all responses as "excellent" and rendering the evaluation ineffective for benchmarking purposes.

Rubric_{llm} suffers from misalignment and unreasonable design. While this baseline demonstrates reasonable discriminative ability with a performance gap of 38.93%, the absolute scores are relatively low (ranging from 28% to 67%), compared to results of the expert-designed rubrics. This pattern suggests that without expert guidance, the LLM-generated criteria may be over rigid or misaligned with the core intent of the research questions. Consequently, the resulting evaluation fails to accurately measure the relevance and quality of the responses.

I.3 Human Evaluation Alignment Results and Analysis

To validate the reliability of different rubric generation methods, we assessed other alignment with human expert judgment using the methodology established in Section 4.3. Domain experts independently annotated a stratified subset of 20 responses using the criteria generated by each method. We then compared these human labels against the automated verdicts provided by the judge model (o3-mini). Table 11 details these alignment metrics — Accuracy, Precision, Recall, and F1-score — for each rubric generation method. The case study is presented in Appendix I.5.

The analysis reveals distinct characteristics for each rubric generation method:

Rubric_{dim} exhibits an artificially inflated F1-score of 0.95. This high score reflects systematic bias rather than genuine alignment. The judge model and human annotators mark nearly all criteria as satisfied, indicating that the generic criteria are excessively lenient and fail to make effective quality distinctions.

Conversely, Rubric_{llm} achieves an F1-score of 0.70 with a notably low Recall of 0.60. This indicates a high rate of false negatives, where the model penalizes responses that human experts consider valid. This pattern suggests that LLM-generated criteria are overly rigid: they often demand specific details present in the reference materials, failing to recognize valid alternative approaches or closely related technologies that effectively address the problem.

In contrast, Expert-Designed rubrics achieve a balanced F1-score of 0.80. This substantial agreement with human experts (Landis & Koch, 1977) demonstrates that expert-designed criteria effectively capture the nuanced judgment required for open-ended AI research evaluation, striking the necessary balance between rigor and flexibility.

I.4 Weight Distribution Analysis

We analyzed the weight distribution patterns across different rubric types to understand how importance assignment differs. Table 12 presents the weight distribution for expert-designed rubrics versus LLM-generated rubrics.

Weight	Expert-Designed	Rubric _{dim}	Interpretation
1 (Nice-to-have)	35.45%	73.66%	Criteria that's beneficial but not essential
2 (Supporting)	50.91%	17.41%	Criteria that add depth and context
3 (Essential)	13.64%	8.93%	Criteria core to understanding the question

Table 12: Weight distribution comparison.

To further analyze the quality of LLM-generated rubrics, we manually assigned weights to Rubric_{dim} following our Weight Assignment Guidelines in Appendix D. According to these guidelines, high weights (3) should be assigned to criteria that are core to understanding the core question, medium weights (2) for supporting criteria that add depth and context, and low weights (1) for nice-to-have criteria that's beneficial but not essential.

After applying these weight assignment standards to Rubric_{dim}, we discovered a problematic distribution pattern: Criteria with weight 1 and 2 dominate (73.66% and 17.41%, respectively), while criteria with weight 3 are severely underrepresented (only 8.93%). This distribution reveals some fundamental limitations in LLM-generated rubric quality:

Insufficient weight 3 (essential criteria): The scarcity of criteria with weight 3 (8.93%) means LLM-generated rubrics fail to emphasize the most critical aspects of open-ended AI research tasks. The deficiency leads to evaluations that cannot effectively distinguish between "adequate" and "excellent" responses, as the rubric fails to prioritize the most important evaluation dimensions. Without sufficient criteria with weight 3, the evaluation cannot focus on what truly matters for answering the research question comprehensively.

Insufficient weight 2 (supporting criteria): The scarcity of criteria with weight 2 (17.41%) indicates a lack of depth assessment. The deficiency in criteria with weight 2 means LLM-generated rubrics largely miss dimensions that assess the depth and comprehensiveness of responses beyond basic requirements. This results in evaluations that cannot distinguish between surface-level and in-depth analyses, as the rubric lacks criteria that assess the supporting details and contextual richness that characterize high-quality research responses.

Over-abundant weight 1 (nice-to-have criteria): The dominance of criteria with weight 1 (73.66%) indicates that LLMs tend to generate "safe" criteria that are easily satisfied but contribute little to distinguishing answer quality. The over-abundance means that evaluation is dominated by peripheral aspects rather than core requirements. This explains the observed ceiling effects: most responses achieve artificially high scores by satisfying

numerous low-importance criteria, masking meaningful differences in actual research quality.

A possible explanation for this phenomenon is that **LLMs lack the domain expertise necessary to identify which aspects of a open-ended AI research task are truly core versus peripheral**. Without deep understanding of AI research priorities and evaluation standards, LLMs default to generating easily satisfiable criteria that avoid making strong claims. This results in a weight distribution that prioritizes peripheral enhancements over core understanding, fundamentally undermining the rubric’s ability to assess research quality effectively.

I.5 Case Studies

To illustrate the qualitative differences between rubric types, we present case studies examining specific evaluation criteria.

I.5.1 Case Study: $Rubric_{dim}$ ’s Lack of Discriminative Power

Case Study 1: Superficial Evaluation Criteria ($Rubric_{dim}$)

Criterion: “Formatting, Layout, and Highlighting for Readability”

Analysis: This criterion focuses entirely on surface-level presentation rather than the substance of the research content. It fails to impose specific requirements related to the research question, allowing models to score highly merely by producing well-formatted text, regardless of whether the actual technical content is accurate or deep.

Case Study 2: Vague and Subjective Criteria ($Rubric_{dim}$)

Criterion: “Originality and Synthesis in the Overall Approach”

Analysis: This criterion relies on abstract qualitative terms without defining specific benchmarks or expected outcomes. It is difficult to objectively quantify “originality” for a research summary task, leading to inconsistent evaluation. The lack of concrete indicators means the judge model often defaults to a positive score as long as the response reads smoothly, failing to distinguish true insight from generic summarization.

I.5.2 Case Study: $Rubric_{llm}$ ’s Over-Strictness

Case Study 3: Overly Rigid Alignment ($Rubric_{llm}$)

Criterion: “Explains how DeepSeek’s combined advances (MoE, MLA, FP8, DualPipe, RL) shift industry trends toward algorithmic/hardware co-design and cost-aware scaling rather than brute-force parameter growth.”

Analysis: This criterion was constructed as a strict checklist of specific terms derived from the reference materials, explicitly requiring the mention of “DualPipe.” Therefore, the evaluation penalized a substantially correct response — which successfully discussed MoE, MLA, FP8, RL, and the core concept of cost efficiency — simply for missing a single keyword.

I.6 Discussion and Conclusions

Our comparative analysis demonstrates that expert-designed rubrics are essential for evaluating open-ended AI research tasks. The key findings are:

1. **LLM-generated rubrics based on generic metrics fail to discriminate.** Rubric_{dim} exhibits severe ceiling effects, with all systems scoring between 87% and 98%, rendering the evaluation ineffective for benchmarking purposes.
2. **LLM-generated rubrics based on reference materials are overly strict and misaligned.** Rubric_{llm} demonstrates better discriminative power but suffers from a low recall of 0.60, indicating that the criteria are overly rigid and fail to recognize valid and equivalent content.
3. **Expert-designed rubrics achieve optimal balance.** Expert-designed rubrics maintain strong discriminative power while achieving a balanced precision-recall profile (F1=0.80), demonstrating substantial alignment with human expert judgment.
4. **Weight distribution reveals fundamental limitations.** LLM-generated rubrics are dominated by peripheral criteria (73.66% criteria with weight 1) and lack sufficient core criteria (8.93% criteria with weight 3). This imbalance indicates that LLMs cannot effectively identify the essential elements of open-ended AI research evaluation.

These findings support our claim that expert-designed rubrics are necessary for reliable evaluation of open-ended AI research questions. The domain expertise required to identify core evaluation dimensions, balance strictness with flexibility, and assign appropriate importance weights cannot be reliably replicated by current LLM-based approaches.

J Score Reliability and Uncertainty Analysis

We present two complementary analyses that together establish the reliability of our scoring: (1) paired bootstrap significance tests to validate that tier-level differences are statistically robust across questions, and (2) a repeated-generation experiment to confirm that within-system generation stochasticity is negligible.

J.1 Paired Bootstrap Significance Tests

To rigorously validate the tier structure reported in Section 5.2, we conducted paired bootstrap significance tests (10,000 resamples, 95% confidence level) between all adjacent system pairs ranked by Coverage in Table 4. For each pair, we resampled the 65 per-question score differences with replacement, computed the mean difference per resample, and derived a one-sided p -value (proportion of resamples where the difference ≤ 0). Effect sizes are reported as Cohen’s d (mean difference / pooled standard deviation).

System Pair	Diff	p -value	Cohen’s d	95% CI
OpenAI vs Gemini	0.010	0.335	0.049	$[-0.037, +0.057]$
Gemini vs Claude	0.082	0.0003	0.375	$[+0.035, +0.128]$
Claude vs Mita	0.028	0.114	0.122	$[-0.018, +0.072]$
Mita vs Doubao	0.008	0.387	0.035	$[-0.043, +0.062]$
Doubao vs Perplexity	0.095	0.0001	0.416	$[+0.052, +0.140]$
Perplexity vs Sonar	0.014	0.234	0.059	$[-0.024, +0.052]$
Sonar vs Grok3 Deeper	0.025	0.130	0.110	$[-0.020, +0.066]$
Grok3 DS vs DeeperS	0.002	0.463	0.007	$[-0.032, +0.033]$
Grok3 DeeperS vs GPT-4o	0.082	0.0002	0.373	$[+0.040, +0.124]$

Table 13: Paired bootstrap significance tests between adjacent system pairs ranked by Coverage. Three statistically significant gaps ($p < 0.001$) define four performance tiers.

Three statistically significant gaps ($p < 0.001$) emerge, defining four performance tiers: {OpenAI Deep Research, Gemini Deep Research} > {Claude Research, Mita Deep Research, Doubao Deep Research} > {Perplexity Deep Research, Perplexity Sonar Reasoning Pro, Grok3 DeepSearch, Grok3 DeeperSearch} > {GPT-4o Search Preview}. Within each tier, differences are not statistically significant ($p > 0.05$), confirming that the fine-grained score differences noted within tiers are indistinguishable from noise. Importantly, the paper’s

main claims are supported by these tests: (1) OpenAI and Gemini Deep Research form a statistically distinct top tier ($p = 0.0003$ vs. the next system); (2) all dedicated DARS in the top three tiers significantly outperform the GPT-4o Search baseline ($p < 0.001$).

J.2 Within-System Generation Stability

To assess whether scores are stable across independent generation runs of each DARS, we conducted a repeated-generation experiment: we sampled a stratified subset of 10 questions and executed 3 independent generation runs for each of the 10 evaluated systems, with all outputs evaluated by o3-mini. Due to the web-UI-based interaction required by most commercial DARS (each query takes 5–15 minutes of manual operation per system), running repeated trials at full benchmark scale is prohibitively time-consuming; we therefore report results on this representative subset.

System	Run 1	Run 2	Run 3	Mean	Std
OpenAI Deep Research	0.7101	0.7083	0.7353	0.7179	± 0.012
Gemini Deep Research	0.6938	0.6924	0.7117	0.6993	± 0.009
Claude Research	0.6202	0.6307	0.6110	0.6206	± 0.008
Mita Deep Research	0.5852	0.5900	0.5984	0.5912	± 0.005
Doubao Deep Research	0.5803	0.5643	0.5951	0.5799	± 0.013
Perplexity Deep Research	0.4773	0.5059	0.4839	0.4890	± 0.012
Perplexity Sonar Reasoning Pro	0.4759	0.4624	0.4792	0.4725	± 0.007
Grok3 DeepSearch	0.4394	0.4486	0.4628	0.4503	± 0.010
Grok3 DeeperSearch	0.4393	0.4326	0.4544	0.4421	± 0.009
GPT-4o Search Preview	0.3629	0.3768	0.3497	0.3631	± 0.011

Table 14: Rubric coverage scores across three independent generation runs on a stratified 10-question subset. The system ranking is identical across all three runs.

The results demonstrate high stability across all systems. The average run-to-run standard deviation is only ± 0.010 , which is $5\times$ **smaller** than the bootstrap CI width (± 0.05) reported in Table 4, and $21\times$ **smaller** than the tier gaps between systems (e.g., Gemini 0.699 vs. Perplexity 0.489, $\Delta = 0.21$). The system ranking is *identical* across all three runs. These findings confirm that run-to-run generation variance is negligible relative to between-system differences, and that DARS stochasticity does not affect system-level rankings.

K Category Balance Analysis

To verify the imbalanced distribution of our dataset does not bias aggregate rankings toward the open consulting performance pattern, we conducted two analyses: (1) equal-weight balanced scoring, re-computing aggregate coverage by assigning equal weight to each question type; and (2) leave-one-category-out analysis, computing rankings after removing each question type in turn.

Analysis	Kendall τ with Original	p -value
Equal-weight balanced (1/3 each)	1.0000	< 0.000001
Leave out Open Consulting	0.9556	0.000006
Leave out Literature Review	0.8667	0.000115
Leave out Technical Details	0.9111	0.000030

Table 15: Rank stability under category rebalancing. All Kendall τ values exceed 0.86 with $p < 0.001$, confirming that system rankings are robust to category distribution.

All Kendall τ values exceed 0.86 with $p < 0.001$, confirming that system rankings remain highly stable under rebalancing and are not artifacts of the category distribution in Table 15.

L Robustness to Length Bias

To examine whether longer responses may naturally cover more rubric items and inflate scores independent of actual research quality, we conduct a controlled length-expansion experiment following the setting of DeepResearch Bench (Du et al., 2025). For every response from Claude Research, we prompt Claude to rewrite it by simply expanding the length while strictly preserving the original information content without introducing new facts, findings, or arguments. We apply this expansion iteratively, yielding two augmented variants whose average article lengths are approximately $\times 1.45$ (*Expanded*) and $\times 2.08$ (*Expanded²*) the original.

DARS	Avg. Length (chars)	Rubric Coverage
Claude Research	15,184	61.13%
Claude Research (Expanded, $\times 1.45$)	21,971	60.91%
Claude Research (Expanded ² , $\times 2.08$)	31,418	59.42%

Table 16: Rubric coverage scores under controlled length expansion of Claude Research responses.

As shown in Table 16, rubric coverage does *not* increase with response length; under the higher expansion factor it actually **declines** from 61.13% to 59.42%. This result shows that our rubric-based evaluation is resilient to simple length-inflation: inflating a response with redundant or padded content cannot boost its score, because each rubric criterion is evaluated independently on whether the corresponding information is substantively present. These findings complement the correlational analysis in Appendix G.7 and together confirm that rubric coverage measures informational substance rather than verbosity.

M Judge Model Sensitivity Analysis

We additionally conducted a cross-judge sensitivity study using five representative systems evaluated by three judge models drawn from distinct model families: **o3-mini** (OpenAI, our primary judge), **Gemini-2.5-Flash** (Google), and **DeepSeek R1** (DeepSeek, open-weight). All judge models followed the same rubric-based scoring protocol described in Section 4.

System	o3-mini	Gemini-2.5-Flash	DeepSeek R1
OpenAI Deep Research	0.7156	0.7174	0.7249
Gemini Deep Research	0.7023	0.7077	0.7079
Perplexity Deep Research	0.4912	0.5047	0.4756
Grok3 DeepSearch	0.4523	0.4556	0.4519
Grok3 DeeperSearch	0.4489	0.4541	0.4428

Table 17: Rubric coverage scores for five systems under three different judge models.

As shown in Table 17, all three judge models produce identical system rankings, and the absolute score differences across judge models are small (within 0.03 for any given system). The consistent ordering, which is OpenAI > Gemini > Perplexity > Grok3 DeepSearch > Grok3 DeeperSearch, holds regardless of whether the judge model is a proprietary OpenAI-series model, a Google-series model, or an open-weight DeepSeek-series model. These results demonstrate that ResearcherBench’s evaluation is robust to judge model selection and does not exhibit vendor or model-family bias.