Novelty Assessment Report

Paper: DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents
PDF URL: https://openreview.net/pdf?id=hQ0K2Hhq7H
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-04

Abstract

Deep Research Agents (DRAs) are emerging as one of the most practical classes of LLM-based agents. Given an open-ended research task, they find, analyze, and synthesize large numbers of online sources to produce a comprehensive report at the level of a research analyst. This can compress hours of manual desk research into minutes. However, a comprehensive benchmark for systematically evaluating the capabilities of these agents remains absent. To bridge this gap, we introduce DeepResearch Bench, a benchmark consisting of 100 PhD-level research tasks, each meticulously crafted by domain experts across 22 distinct fields. To evaluate DRAs comprehensively, we propose two complementary and fully automated methodologies. The first is a reference-based method with adaptive criteria to assess the quality of generated research reports. The second evaluates a DRA's information-retrieval and collection capabilities by assessing its effective citation count and overall citation accuracy. By conducting extensive human consistency experiments, we demonstrate that our evaluation methods are highly aligned with expert judges and faithfully reflect human judgments of quality differences among DRA-generated content. We are open-sourcing DeepResearch Bench and key components of these frameworks to accelerate the development of practical LLM-based agents.

Core Task Landscape

This paper addresses: **Evaluating Deep Research Agents**

A total of **50 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Benchmark Design and Construction**
- **Evaluation Methodologies**
- **Agent Training and Optimization**
- **Agent Architectures and Systems**
- **Conceptual Foundations and Surveys**
- **Improving Existing Systems**
- **Non-Research Agent Deep Learning**

Complete Taxonomy Tree

- Evaluating Deep Research Agents Survey Taxonomy
- Benchmark Design and Construction
  - General Deep Research Benchmarks ★ (6 papers)
  - [0] DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents (Anon et al., 2026) View paper
  - [2] Deep Research Bench: Evaluating AI Web Research Agents (- -, 2025) View paper
  - [7] Characterizing deep research: A benchmark and formal definition (Java, 2025) View paper
  - [8] ResearchRubrics: A Benchmark of Prompts and Rubrics For Evaluating Deep Research Agents (Manasi Sharma, 2025) View paper
  - [13] Liveresearchbench: A live benchmark for user-centric deep research in the wild (Wang Jia-Yu, 2025) View paper
  - [16] Reportbench: Evaluating deep research agents via academic survey tasks (Li Ming-hao, 2025) View paper
  - Domain-Specific Benchmarks (6 papers)
  - [3] MedBrowseComp: Benchmarking Medical Deep Research and Computer Use (Chen Shan, 2025) View paper
  - [5] Researcherbench: Evaluating deep ai research systems on the frontiers of scientific inquiry (Xu, 2025) View paper
  - [36] DeepResearch Arena: The First Exam of LLMs' Research Abilities via Seminar-Grounded Tasks (Wan Hai-yuan, 2025) View paper
  - [38] DRBench: A Realistic Benchmark for Enterprise Deep Research (Abaskohi, 2025) View paper
  - [40] FinResearchBench: A Logic Tree based Agent-as-a-Judge Evaluation Framework for Financial Research Agents (Sun Rui, 2025) View paper
  - [48] FinDeepResearch: Evaluating Deep Research Agents in Rigorous Financial Analysis (Zhu Fengbin, 2025) View paper
  - Machine Learning Research Task Benchmarks (7 papers)
  - [9] Mlgym: A new framework and benchmark for advancing ai research agents (Nathani, 2025) View paper
  - [12] Benchmarking large language models as ai research agents (Huang Qian, 2023) View paper
  - [20] Mlagentbench: Evaluating language agents on machine learning experimentation (Huang Qian, 2023) View paper
  - [25] AI Agents for Deep Scientific Research (R Zhou, 2025) View paper
  - [28] MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research (Chen Hui, 2025) View paper
  - [29] Astabench: Rigorous benchmarking of ai agents with a scientific research suite (Bragg, 2025) View paper

- [41] ResearchArena: Benchmarking Large Language Models' Ability to Collect and Organize Information as Research Agents (Hao Kang, 2024) View paper
- Specialized Application Benchmarks (2 papers)
- [11] DeepShop: A Benchmark for Deep Research Shopping Agents (Lyu, 2025) View paper
- [21] Mind2Web 2: Evaluating Agentic Search with Agent-as-a-Judge (Huang, 2025) View paper

- Evaluation Methodologies
  - Automated Evaluation Frameworks (2 papers)
  - [18] Mcpeval: Automatic mcp-based deep evaluation for ai agent models (Liu ZhiWei, 2025) View paper
  - [26] AI Research Agents for Machine Learning: Search, Exploration, and Generalization in MLE-bench (Hambardzumyan, 2025) View paper
  - Human Evaluation Platforms (1 papers)
  - [23] Deep research comparator: A platform for fine-grained human annotations of deep research agents (Jin Jia-he, 2025) View paper
- Agent Training and Optimization
  - Reinforcement Learning for Research Agents (4 papers)
  - [1] Deepresearcher: Scaling deep research via reinforcement learning in real-world environments (Zheng Yu-Xiang, 2025) View paper
  - [6] Reinforcement learning foundations for deep research systems: A survey (Li Wen-Jun, 2025) View paper
  - [17] PoU: Proof-of-Use to Counter Tool-Call Hacking in DeepResearch Agents (Ma Shengjie, 2025) View paper
  - [30] Sfr-deepresearch: Towards effective reinforcement learning for autonomously reasoning single agents (Nguyen, 2025) View paper
  - Alternative Training Paradigms (2 papers)
  - [14] Tongyi deepresearch technical report (Tongyi Li, 2025) View paper
  - [33] Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework (Yu Ai-ling, 2025) View paper
- Agent Architectures and Systems
  - Single-Agent Research Systems (3 papers)
  - [19] WebWatcher: Breaking New Frontier of Vision-Language Deep Research Agent (Geng Xin-yu, 2025) View paper
  - [27] Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents (Qiao, 2025) View paper
  - [42] WebWeaver: Structuring Web-Scale Evidence with Dynamic Outlines for Open-Ended Deep Research (Li Zijian, 2025) View paper
  - Multi-Agent Research Systems (2 papers)
  - [32] The Denario project: Deep knowledge AI agents for scientific discovery (Villaescusa-Navarro, 2025) View paper
  - [44] Agentrxiv: Towards collaborative autonomous research (Schmidgall, 2025) View paper
  - Specialized Research Agent Applications (3 papers)
  - [10] Mlr-copilot: Autonomous machine learning research based on large language models agents (Li Ruochen, 2024) View paper
  - [34] Geak: Introducing Triton Kernel AI Agent & Evaluation Benchmarks (Wang Jianghui, 2025) View paper
  - [47] AI-Researcher: Autonomous Scientific Innovation (Tang, 2025) View paper
- Conceptual Foundations and Surveys (3 papers)
  - [4] Deep research: A survey of autonomous research agents (Zhang Wen-lin, 2025) View paper
  - [22] From Web Search towards Agentic Deep Research: Incentivizing Search with Reasoning Agents (Zhang Weizhi, 2025) View paper
  - [24] A survey of llm-based deep search agents: Paradigm, optimization, evaluation, and challenges (Xi, 2025) View paper
- Improving Existing Systems (2 papers)
  - [15] Improving and Evaluating Open Deep Research Agents (Bradbury, 2025) View paper
  - [35] What Does It Take to Be a Good AI Research Agent? Studying the Role of Ideation Diversity (Alexis Audran-Reiss, 2025) View paper
- Non-Research Agent Deep Learning (8 papers)
  - [31] Study of Q-learning and deep Q-network learning control for a rotary inverted pendulum system (Zied Ben Hazem, 2024) View paper
  - [37] An Adaptive Agent Decision Model Based on Deep Reinforcement Learning and Autonomous Learning (Zhu, 2023) View paper
  - [39] Deep Reinforcement Learning-Based Multi-Agent System with Advanced Actorâ□□Critic Framework for Complex Environment (Zihao Cui, 2025) View paper
  - [43] From Deep Learning to LLMs: A survey of AI in Quantitative Investment (Cao Bokai, 2025) View paper
  - [45] Multi-Agent Deep Reinforcement Learning for Coordinated Energy Trading and Flexibility Services Provision in Local Electricity Markets (Yujian Ye, 2023) View paper
  - [46] Emergent multi-agent communication in the deep learning era (Lazaridou, 2020) View paper
  - [49] Collaborative Multi-Agent Deep Reinforcement Learning for Energy-Efficient Resource Allocation in Heterogeneous Mobile Edge Computing Networks (Yang Xiao, 2024) View paper
  - [50] Deep learning-based natural language processing in humanâ□□agent interaction: Applications, advancements and challenges (Nafiz Ahmed, 2024) View paper

Narrative

Core task: Evaluating deep research agents. The field has organized itself around several complementary dimensions. Benchmark Design and Construction focuses on creating standardized testbeds that capture the complexity of research tasks, ranging from general deep research challenges to domain-specific scenarios in medicine, finance, and machine learning. Evaluation Methodologies develops principled frameworks for assessing agent capabilities, including rubric-based scoring and comparative analysis approaches. Agent Training and Optimization explores reinforcement learning and other techniques to improve agent performance, while Agent Architectures and Systems examines the structural designs that enable effective research behavior. Conceptual Foundations and Surveys provide theoretical grounding and landscape overviews, such as Deep Research Survey[4] and Characterizing Deep Research[7]. Improving Existing Systems targets incremental enhancements to deployed agents, and Non-Research Agent Deep Learning addresses related but distinct applications of deep learning in agentic contexts.

Within Benchmark Design and Construction, a particularly active cluster has emerged around general deep research benchmarks that attempt to capture the full scope of research activities—from literature review and hypothesis generation to experimental design and report writing. Works like Deep Research Bench[2], ResearcherBench[5], and LiveResearchBench[13] each propose different task

formulations and evaluation protocols, reflecting ongoing debates about what constitutes a faithful representation of research work. DeepResearch Bench[0] situates itself squarely in this general benchmark cluster, emphasizing comprehensive evaluation across multiple research stages. Compared to ResearcherBench[5], which may focus on particular research subtasks, and LiveResearchBench[13], which incorporates dynamic or real-time elements, DeepResearch Bench[0] appears to prioritize breadth and standardization in capturing the research process, contributing another perspective to the evolving question of how best to measure deep research agent capabilities.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. Deep Research Bench: Evaluating AI Web Research Agents

Authors: - -, Nikos I. Bosse, Bosse, Nikos I., Jon Evans, et al. (16 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Amongst the most common use cases of modern AI is LLM chat with web search enabled. However, no direct evaluations of the quality of web research agents exist that control for the continually-changing web. We introduce Deep Research Bench, consisting of 89 multi-step web research task instances of varying difficulty across 8 diverse task categories, with the answers carefully worked out by skilled humans. We provide a"RetroSearch"environment with a large frozen set of scraped web pages, and demo...

Relationship Analysis

Both papers belong to the General Deep Research Benchmarks category, focusing on evaluating AI agents that perform multi-step web research tasks across diverse domains. They overlap in their core objective of benchmarking deep research agents through comprehensive task sets (100 vs 89 tasks) and automated evaluation methodologies. The key differences are that the original paper (DeepResearch Bench) emphasizes PhD-level tasks with domain expert curation, introduces the RACE and FACT evaluation frameworks for report quality and citation accuracy, and derives task distribution from 96K real user queries, while the candidate paper (Deep Research Bench) focuses on a RetroSearch environment with frozen web pages to enable reproducible evaluations over time and includes explicit evaluation of commercial deep research products.

---

2. Characterizing deep research: A benchmark and formal definition

Authors: Java, Abhinav, Abhinav Java, Ashmit Khandelwal, Halfaker, et al. (22 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Information tasks such as writing surveys or analytical reports require complex search and reasoning, and have recently been grouped under the umbrella of \textit{deep research} -- a term also adopted by recent models targeting these capabilities. Despite growing interest, the scope of the deep research task remains underdefined and its distinction from other reasoning-intensive problems is poorly understood. In this paper, we propose a formal characterization of the deep research (DR) task and ...

Relationship Analysis

Both papers belong to the General Deep Research Benchmarks category, focusing on creating comprehensive evaluation frameworks for deep research agents across diverse domains without domain-specific constraints. They overlap in their core objective of systematically evaluating deep research agents through multi-domain benchmarks (DeepResearchBench with 100 PhD-level tasks across 22 fields, LiveDRBench with 100 tasks across scientific and world events domains) and both propose automated evaluation methodologies. The key differences are that DeepResearchBench emphasizes report quality evaluation through reference-based adaptive criteria (RACE) and citation accuracy (FACT), while LiveDRBench focuses on formal problem characterization through claim-based evaluation and introduces a problem inversion methodology to create tasks that avoid single-document answers on the web.

---

3. ResearchRubrics: A Benchmark of Prompts and Rubrics For Evaluating Deep Research Agents

Authors: Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, et al. (16 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Deep Research (DR) is an emerging agent application that leverages large language models (LLMs) to address open-ended queries. It requires the integration of several capabilities, including multi-step reasoning, cross-document synthesis, and the generation of evidence-backed, long-form answers. Evaluating DR remains challenging because responses are lengthy and diverse, admit many valid solutions, and often depend on dynamic information sources. We introduce ResearchRubrics, a standardized bench...

Relationship Analysis

Both papers belong to the General Deep Research Benchmarks category, focusing on evaluating deep research agents across diverse open-ended tasks without domain-specific constraints. They overlap in their goal of providing comprehensive benchmarks with expert-crafted tasks and automated evaluation frameworks for assessing report quality and agent capabilities. The key differences are that DeepResearch Bench emphasizes a data-driven approach with 100 PhD-level tasks derived from 96K user queries and introduces RACE/FACT evaluation frameworks, while ResearchRubrics focuses on 2,500+ expert-written fine-grained rubrics with a complexity framework categorizing tasks along conceptual breadth, logical nesting, and exploration axes.

---

4. Liveresearchbench: A live benchmark for user-centric deep research in the wild

Authors: Wang Jia-Yu, Ming, Yifei, Jiayu Wang, Yifei Ming, et al. (25 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Deep research -- producing comprehensive, citation-grounded reports by searching and synthesizing information from hundreds of live web sources -- marks an important frontier for agentic systems. To rigorously evaluate this ability, four principles are essential: tasks should be (1) user-centric, reflecting realistic information needs, (2) dynamic, requiring up-to-date information beyond parametric knowledge, (3) unambiguous, ensuring consistent interpretation across users, and (4) multi-faceted...

Relationship Analysis

Both papers belong to the General Deep Research Benchmarks category, focusing on evaluating deep research agents through comprehensive benchmarks covering diverse open-ended research tasks. They overlap in their core objective of assessing agents' abilities to search, synthesize, and generate research reports, and both propose automated evaluation frameworks for report quality and citation accuracy. However, DeepResearchBench emphasizes PhD-level tasks derived from 96K user queries with reference-based adaptive criteria (RACE), while LiveResearchBench prioritizes user-centric, unambiguous, and time-varying tasks with a comprehensive DeepEval suite covering six complementary dimensions including presentation, consistency, and depth analysis.

---

5. Reportbench: Evaluating deep research agents via academic survey tasks

Authors: Li Ming-hao, Zeng Ying, Minghao Li, Cheng, Zhihao, et al. (11 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The advent of Deep Research agents has substantially reduced the time required for conducting extensive research tasks. However, these tasks inherently demand rigorous standards of factual accuracy and comprehensiveness, necessitating thorough evaluation before widespread adoption. In this paper, we propose ReportBench, a systematic benchmark designed to evaluate the content quality of research reports generated by large language models (LLMs). Our evaluation focuses on two critical dimensions: ...

Relationship Analysis

Both papers belong to the General Deep Research Benchmarks category, focusing on evaluating deep research agents through comprehensive benchmark construction and automated evaluation methodologies. They overlap in addressing the challenge of assessing research report quality and citation accuracy for deep research agents. The key difference is that DeepResearch Bench uses 100 PhD-level tasks across 22 domains derived from 96K user queries with the RACE and FACT evaluation frameworks, while ReportBench specifically leverages published arXiv survey papers as gold-standard references and focuses on academic survey tasks with citation faithfulness verification.

Contributions Analysis

Overall novelty summary. The paper introduces DeepResearch Bench, a benchmark comprising 100 PhD-level research tasks across 22 fields, alongside two automated evaluation methodologies (RACE for report quality, FACT for citation accuracy). It resides in the 'General Deep Research Benchmarks' leaf, which contains six papers total, indicating a moderately populated research direction. This leaf sits within the broader 'Benchmark Design and Construction' branch, suggesting the paper contributes to an active area focused on standardizing evaluation infrastructure for deep research agents.

The taxonomy reveals neighboring leaves addressing domain-specific benchmarks (medicine, finance, scientific research) and machine learning experimentation tasks, as well as a separate branch for evaluation methodologies. DeepResearch Bench bridges benchmark construction and evaluation methods by proposing both a dataset and assessment frameworks. Its emphasis on general-purpose, cross-domain tasks distinguishes it from domain-specific benchmarks, while its automated evaluation approach connects to the 'Automated Evaluation Frameworks' leaf under 'Evaluation Methodologies,' though it remains classified primarily as a benchmark contribution.

Among 30 candidates examined, none clearly refute the three core contributions. The DeepResearch Bench dataset examined 10 candidates with zero refutable overlaps; similarly, the RACE and FACT evaluation frameworks each examined 10 candidates with no refutations. This suggests that within the limited search scope, the specific combination of PhD-level task design, adaptive reference-based evaluation, and citation-accuracy metrics appears relatively novel. However, the presence of five sibling papers in the same taxonomy leaf indicates that general deep research benchmarking is an established direction with existing proposals.

Based on the top-30 semantic matches and taxonomy structure, the work appears to offer a distinct contribution to a moderately crowded benchmark landscape. The lack of refutable overlaps across all three contributions within this limited scope suggests differentiation from prior work, though the analysis does not cover exhaustive literature or adjacent evaluation methodology papers outside the examined candidates. The taxonomy context indicates the paper extends an active research thread rather than opening an entirely new direction.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: DeepResearch Bench benchmark dataset

Description: A specialized benchmark for evaluating Deep Research Agents, constructed through large-scale analysis of over 96,000 real user queries and expert collaboration. The benchmark contains 100 tasks across 22 domains, designed to balance challenge while reflecting authentic user needs.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization

URL: View paper

Brief Assessment

QMSum[71] focuses on query-based meeting summarization across product, academic, and committee meetings, not PhD-level research tasks across academic domains. The datasets serve fundamentally different purposes and task types.

---

2. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition

URL: View paper

Brief Assessment

ResearchBench[70] focuses on scientific hypothesis discovery tasks (inspiration retrieval, hypothesis composition, hypothesis ranking) across 12 scientific disciplines, not on evaluating deep research agents that produce comprehensive research reports from open-ended queries across 22 domains.

---

3. MSEarth: A Multimodal Scientific Dataset and Benchmark for Phenomena Uncovering in Earth Science

URL: View paper

Brief Assessment

MSEarth[74] focuses on multimodal scientific reasoning in Earth science with figure-caption pairs, not PhD-level research task evaluation across multiple domains like DeepResearch Bench.

---

4. MMSci: A Dataset for Graduate-Level Multi-Discipline Multimodal Scientific Understanding

URL: View paper

Brief Assessment

MMSci[72] focuses on scientific figure interpretation from Nature Communications articles across 72 fields, not on evaluating deep research agents or PhD-level research task completion across domains.

---

5. Making the implicit explicit: Creating performance expectations for the dissertation

URL: View paper

Brief Assessment

Dissertation Performance Expectations[69] focuses on creating explicit performance standards for doctoral dissertations across 10 academic disciplines through faculty focus groups. It does not address AI research agents, benchmark construction for evaluating automated systems, or PhD-level research task evaluation in the context of machine learning agents.

---

6. DeepResearch Arena: The First Exam of LLMs' Research Abilities via Seminar-Grounded Tasks

URL: View paper

Brief Assessment

DeepResearch Arena[36] focuses on seminar-grounded tasks extracted from academic discourse, while the original paper constructs tasks through expert collaboration based on user query distribution analysis. The data sources and construction methodologies are fundamentally different.

---

7. Scicode: A research coding benchmark curated by scientists

URL: View paper

Brief Assessment

SciCode[68] focuses on research coding tasks in natural sciences (physics, chemistry, biology), not on evaluating deep research agents that synthesize online sources into comprehensive reports. The benchmarks address fundamentally different capabilities.

---

8. SentiGrad: A New Hindi-English Code Mixed Sentiment Analysis Dataset with Preliminary Results and Open Challenges

URL: View paper

Brief Assessment

SentiGrad[73] focuses on Hindi-English code-mixed sentiment analysis for educational content comments, not PhD-level research task benchmarks across academic domains. The datasets serve entirely different purposes and domains.

---

9. Towards Personalized Deep Research: Benchmarks and Evaluations

URL: View paper

Brief Assessment

Personalized Deep Research[67] focuses on personalization in deep research agents with user profiles and persona-driven adaptation, while the original paper emphasizes PhD-level task complexity across domains without personalization dimensions. These are complementary evaluation perspectives rather than overlapping novelty claims.

---

10. ResearchRubrics: A Benchmark of Prompts and Rubrics For Evaluating Deep Research Agents

URL: View paper

Brief Assessment

ResearchRubrics[8] focuses on evaluation methodology (rubrics and protocols) rather than benchmark construction. While both address deep research evaluation, ResearchRubrics[8] does not demonstrate prior work on constructing PhD-level task benchmarks from large-scale user query analysis across 22 domains.

Contribution 2: RACE evaluation framework

Description: A Reference-based and Adaptive Criteria-driven Evaluation framework with Dynamic Weighting that assesses research report quality. The framework dynamically generates task-specific weights and criteria, then employs reference-based scoring to evaluate reports across four dimensions: Comprehensiveness, Insight/Depth, Instruction-Following, and Readability.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. FinRpt: Dataset, Evaluation System and LLM-based Multi-agent Framework for Equity Research Report Generation

URL: View paper

Brief Assessment

FinRpt[66] focuses on equity research report generation with LLM-based evaluation metrics for financial professionalism (financial numeric, news analysis, company/market/industry insights, etc.), not on general research report quality assessment with dynamic criteria generation and reference-based scoring as in RACE.

---

2. Earnings2Insights: Analyst Report Generation for Investment Guidance

URL: View paper

Brief Assessment

Earnings2Insights[62] focuses on evaluating investment report generation using decision-based human evaluation and LLM-as-a-judge for pairwise comparisons. It does not propose a reference-based adaptive criteria-driven framework with dynamic weighting for research report quality assessment.

---

3. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards

URL: View paper

Brief Assessment

Radiology Semantic Rewards[63] focuses on radiology report generation from chest X-rays using semantic graph-based rewards (RadGraph), not on evaluating general research reports across multiple domains with dynamic criteria generation.

---

4. Knowledge Distillation and Transformer-Based Framework for Automatic Spine CT Report Generation

URL: View paper

Brief Assessment

Spine Report Generation[59] focuses on automated medical report generation for spine CT scans using transformer architectures and knowledge distillation. It does not propose or discuss any reference-based evaluation framework for assessing research report quality.

---

5. Researcherbench: Evaluating deep ai research systems on the frontiers of scientific inquiry

URL: View paper

Brief Assessment

ResearcherBench[5] focuses on evaluating deep AI research systems on frontier scientific questions using expert-designed rubrics for insight quality, not general research report quality assessment with dynamic criteria generation.

---

6. Reportbench: Evaluating deep research agents via academic survey tasks

URL: View paper

Brief Assessment

ReportBench[16] focuses on evaluating research reports using published survey papers as references and checking citation faithfulness, rather than proposing a general reference-based adaptive criteria framework with dynamic weighting for report quality assessment across multiple dimensions.

---

7. YUNet_LLMClaimReport: An Enhanced Automobile Insurance Fraud Detection and Automated Claim Report Generation Using Large Language Models

URL: View paper

Brief Assessment

Insurance Fraud Detection[61] focuses on automobile insurance fraud detection and automated claim report generation using computer vision and LLMs. It does not address reference-based evaluation frameworks for research report quality assessment, which is the core contribution of the original paper's RACE framework.

---

8. Multimodal DeepResearcher: Generating Text-Chart Interleaved Reports From Scratch with Agentic Framework

URL: View paper

Brief Assessment

Multimodal DeepResearcher[64] focuses on generating text-chart interleaved reports with visualization integration, not on reference-based evaluation frameworks for assessing research report quality. The candidate's evaluation component (MultimodalReportBench) serves a different purpose than RACE's adaptive criteria-driven assessment methodology.

---

9. Slit Lamp Report Generation and Question Answering: Development and Validation of a Multimodal Transformer Model with Large Language Model Integration

URL: View paper

Brief Assessment

Slit Lamp Report[65] focuses on medical image report generation using qualitative metrics (BLEU, CIDEr, ROUGE-L, SPICE) and ophthalmologist assessments, not on reference-based evaluation frameworks for research report quality with dynamic weighting and adaptive criteria generation.

---

10. On the Evaluation of Machine-Generated Reports

URL: View paper

Brief Assessment

Machine Generated Reports[60] focuses on evaluating reports through nugget-based content assessment and citation verification, not on reference-based adaptive criteria with dynamic weighting as in RACE.

Contribution 3: FACT evaluation framework

Description: A framework for Factual Abundance and Citation Trustworthiness that evaluates Deep Research Agents' information-retrieval and collection capabilities by assessing effective citation count and overall citation accuracy.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Application of Generative Artificial Intelligence Models for Accurate Prescription Label Identification and Information Retrieval for the Elderly in Northern East of Thailand

URL: View paper

Brief Assessment

Prescription Label AI[53] focuses on medication label interpretation for elderly patients using RAGAS metrics (context recall, factual correctness, faithfulness, semantic similarity) to evaluate AI model performance. This differs fundamentally from FACT, which evaluates Deep Research Agents through citation accuracy and effective citation counts for web-based information retrieval tasks.

---

2. TrustRAG: An Information Assistant with Retrieval Augmented Generation

URL: View paper

Brief Assessment

TrustRAG[58] focuses on enhancing RAG system trustworthiness through semantic chunking, utility-based filtering, and fine-grained citation at the sentence level for excerpt-based QA tasks. The ORIGINAL paper's FACT framework evaluates Deep Research Agents by measuring citation accuracy and effective citation counts across comprehensive research reports, which is a different evaluation context and methodology.

---

3. Researcherbench: Evaluating deep ai research systems on the frontiers of scientific inquiry

URL: View paper

Brief Assessment

ResearcherBench[5] introduces faithfulness and groundedness metrics for factual assessment, but these evaluate citation accuracy differently from FACT's effective citation count and overall citation accuracy metrics.

---

4. Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy

URL: View paper

Brief Assessment

Health Information RAG[57] focuses on evaluating retrieved health documents through factual accuracy and topical relevance metrics for information retrieval systems, not on evaluating Deep Research Agents' citation capabilities or effective citation counts as proposed in the original paper's FACT framework.

---

5. PoU: Proof-of-Use to Counter Tool-Call Hacking in DeepResearch Agents

URL: View paper

Brief Assessment

Proof of Use[17] focuses on training RL agents to ground reasoning in retrieved evidence through citation contracts and perturbation rewards, not on evaluating information retrieval systems. The candidate addresses agent training methodology, while the original contribution concerns benchmark evaluation metrics for Deep Research Agents.

---

6. AI chatbot accountability in the age of algorithmic gatekeeping: Comparing generative search engine political information retrieval across five languages

URL: View paper

Brief Assessment

Chatbot Political Retrieval[54] focuses on evaluating AI chatbot responses for political information retrieval across languages, examining factual correctness and attribution behaviors in a specific political context. This differs from FACT's framework for evaluating Deep Research Agents' citation accuracy and effective citation counts across general research tasks.

---

7. Investigations on using Evidence-Based GraphRag Pipeline using LLM Tailored for Answering USMLE Medical Exam Questions

URL: View paper

Brief Assessment

GraphRAG USMLE[55] focuses on citation fidelity in medical QA systems using knowledge graphs, not on evaluating deep research agents' information-retrieval capabilities through factual abundance metrics.

---

8. Integrating Information Retrieval and LLMs: A Document Retrieval Chatbot in Education Settings

URL: View paper

Brief Assessment

Document Retrieval Chatbot[56] focuses on educational chatbot systems with basic retrieval accuracy metrics, not comprehensive deep research agent evaluation frameworks with citation trustworthiness and factual abundance assessment.

---

9. Statistical biases in information retrieval metrics for recommender systems

URL: View paper

Brief Assessment

Recommender Metrics Bias[51] focuses on statistical biases in information retrieval metrics for recommender systems evaluation, specifically addressing sparsity and popularity biases in metrics like precision and MAP. This is fundamentally different from FACT, which evaluates Deep Research Agents' citation accuracy and factual abundance in research report generation tasks.

---

10. CiteFix: Enhancing RAG Accuracy Through Post-Processing Citation Correction

URL: View paper

Brief Assessment

CiteFix[52] focuses on post-processing citation correction in RAG systems to improve accuracy, not on evaluating deep research agents' information-retrieval capabilities through citation metrics as a benchmark framework.

Appendix: Text Similarity Detection

Textual similarity detection checked 32 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Multimodal DeepResearcher: Generating Text-Chart Interleaved Reports From Scratch with Agentic Framework

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents View paper
- [1] Deepresearcher: Scaling deep research via reinforcement learning in real-world environments View paper
- [2] Deep Research Bench: Evaluating AI Web Research Agents View paper
- [3] MedBrowseComp: Benchmarking Medical Deep Research and Computer Use View paper
- [4] Deep research: A survey of autonomous research agents View paper
- [5] Researcherbench: Evaluating deep ai research systems on the frontiers of scientific inquiry View paper
- [6] Reinforcement learning foundations for deep research systems: A survey View paper
- [7] Characterizing deep research: A benchmark and formal definition View paper
- [8] ResearchRubrics: A Benchmark of Prompts and Rubrics For Evaluating Deep Research Agents View paper
- [9] Mlgym: A new framework and benchmark for advancing ai research agents View paper
- [10] Mlr-copilot: Autonomous machine learning research based on large language models agents View paper
- [11] DeepShop: A Benchmark for Deep Research Shopping Agents View paper
- [12] Benchmarking large language models as ai research agents View paper
- [13] Liveresearchbench: A live benchmark for user-centric deep research in the wild View paper
- [14] Tongyi deepresearch technical report View paper
- [15] Improving and Evaluating Open Deep Research Agents View paper
- [16] Reportbench: Evaluating deep research agents via academic survey tasks View paper
- [17] PoU: Proof-of-Use to Counter Tool-Call Hacking in DeepResearch Agents View paper
- [18] Mcpeval: Automatic mcp-based deep evaluation for ai agent models View paper
- [19] WebWatcher: Breaking New Frontier of Vision-Language Deep Research Agent View paper
- [20] Mlagentbench: Evaluating language agents on machine learning experimentation View paper
- [21] Mind2Web 2: Evaluating Agentic Search with Agent-as-a-Judge View paper
- [22] From Web Search towards Agentic Deep Research: Incentivizing Search with Reasoning Agents View paper
- [23] Deep research comparator: A platform for fine-grained human annotations of deep research agents View paper
- [24] A survey of llm-based deep search agents: Paradigm, optimization, evaluation, and challenges View paper
- [25] AI Agents for Deep Scientific Research View paper
- [26] AI Research Agents for Machine Learning: Search, Exploration, and Generalization in MLE-bench View paper
- [27] Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents View paper
- [28] MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research View paper
- [29] Astabench: Rigorous benchmarking of ai agents with a scientific research suite View paper
- [30] Sfr-deepresearch: Towards effective reinforcement learning for autonomously reasoning single agents View paper

- [31] Study of Q-learning and deep Q-network learning control for a rotary inverted pendulum system View paper
- [32] The Denario project: Deep knowledge AI agents for scientific discovery View paper
- [33] Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework View paper
- [34] Geak: Introducing Triton Kernel AI Agent & Evaluation Benchmarks View paper
- [35] What Does It Take to Be a Good AI Research Agent? Studying the Role of Ideation Diversity View paper
- [36] DeepResearch Arena: The First Exam of LLMs' Research Abilities via Seminar-Grounded Tasks View paper
- [37] An Adaptive Agent Decision Model Based on Deep Reinforcement Learning and Autonomous Learning View paper
- [38] DRBench: A Realistic Benchmark for Enterprise Deep Research View paper
- [39] Deep Reinforcement Learning-Based Multi-Agent System with Advanced Actorâ□□Critic Framework for Complex Environment View paper
- [40] FinResearchBench: A Logic Tree based Agent-as-a-Judge Evaluation Framework for Financial Research Agents View paper
- [41] ResearchArena: Benchmarking Large Language Models' Ability to Collect and Organize Information as Research Agents View paper
- [42] WebWeaver: Structuring Web-Scale Evidence with Dynamic Outlines for Open-Ended Deep Research View paper
- [43] From Deep Learning to LLMs: A survey of AI in Quantitative Investment View paper
- [44] Agentrxiv: Towards collaborative autonomous research View paper
- [45] Multi-Agent Deep Reinforcement Learning for Coordinated Energy Trading and Flexibility Services Provision in Local Electricity Markets View paper
- [46] Emergent multi-agent communication in the deep learning era View paper
- [47] AI-Researcher: Autonomous Scientific Innovation View paper
- [48] FinDeepResearch: Evaluating Deep Research Agents in Rigorous Financial Analysis View paper
- [49] Collaborative Multi-Agent Deep Reinforcement Learning for Energy-Efficient Resource Allocation in Heterogeneous Mobile Edge Computing Networks View paper
- [50] Deep learning-based natural language processing in humanâ□□agent interaction: Applications, advancements and challenges View paper
- [51] Statistical biases in information retrieval metrics for recommender systems View paper
- [52] CiteFix: Enhancing RAG Accuracy Through Post-Processing Citation Correction View paper
- [53] Application of Generative Artificial Intelligence Models for Accurate Prescription Label Identification and Information Retrieval for the Elderly in Northern East of Thailand View paper
- [54] AI chatbot accountability in the age of algorithmic gatekeeping: Comparing generative search engine political information retrieval across five languages View paper
- [55] Investigations on using Evidence-Based GraphRag Pipeline using LLM Tailored for Answering USMLE Medical Exam Questions View paper
- [56] Integrating Information Retrieval and LLMs: A Document Retrieval Chatbot in Education Settings View paper
- [57] Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy View paper
- [58] TrustRAG: An Information Assistant with Retrieval Augmented Generation View paper
- [59] Knowledge Distillation and Transformer-Based Framework for Automatic Spine CT Report Generation View paper
- [60] On the Evaluation of Machine-Generated Reports View paper
- [61] YUNet_LLMClaimReport: An Enhanced Automobile Insurance Fraud Detection and Automated Claim Report Generation Using Large Language Models View paper
- [62] Earnings2Insights: Analyst Report Generation for Investment Guidance View paper
- [63] Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards View paper
- [64] Multimodal DeepResearcher: Generating Text-Chart Interleaved Reports From Scratch with Agentic Framework View paper
- [65] Slit Lamp Report Generation and Question Answering: Development and Validation of a Multimodal Transformer Model with Large Language Model Integration View paper
- [66] FinRpt: Dataset, Evaluation System and LLM-based Multi-agent Framework for Equity Research Report Generation View paper
- [67] Towards Personalized Deep Research: Benchmarks and Evaluations View paper
- [68] Scicode: A research coding benchmark curated by scientists View paper
- [69] Making the implicit explicit: Creating performance expectations for the dissertation View paper
- [70] Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition View paper
- [71] QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization View paper
- [72] MMSci: A Dataset for Graduate-Level Multi-Discipline Multimodal Scientific Understanding View paper
- [73] SentiGrad: A New Hindi-English Code Mixed Sentiment Analysis Dataset with Preliminary Results and Open Challenges View paper
- [74] MSEarth: A Multimodal Scientific Dataset and Benchmark for Phenomena Uncovering in Earth Science View paper