Novelty Assessment Report

Paper: AstaBench: Rigorous Benchmarking of AI Agents with a Scientific Research Suite
PDF URL: https://openreview.net/pdf?id=M7TNf5J26u
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

AI agents hold the potential to revolutionize scientific productivity by automating literature reviews, replicating experiments, analyzing data, and even proposing new directions of inquiry; indeed, there are now many such agents, ranging from general-purpose "deep research" systems to specialized science-specific agents, such as AI Scientist and AIGS. Rigorous evaluation of these agents is critical for progress. Yet existing benchmarks fall short on several fronts: they often (1) lack reproducible agent tools necessary for a controlled comparison of core agentic capabilities; (2) do not account for confounding variables such as model cost and tool access; (3) do not provide standardized interfaces for quick agent prototyping and evaluation; (4) fail to provide holistic, product-informed measures of real-world use cases such as science research; and (5) lack comprehensive baseline agents necessary to identify true advances. In response, we define principles and tooling for more rigorously benchmarking agents. Using these, we present AstaBench, a suite that provides a holistic measure of agentic ability to perform scientific research, comprising 2400+ problems spanning the entire scientific discovery process and multiple scientific domains, and including many problems inspired by actual user requests to deployed Asta agents. Our suite comes with the first scientific research environment with production-grade search tools that enable controlled, reproducible evaluation, better accounting for confounders. Alongside, we provide a comprehensive suite of nine science-optimized classes of Asta agents and numerous baselines. Our extensive evaluation of 57 agents across 22 agent classes reveals several interesting findings, most importantly that despite meaningful progress on certain individual aspects, AI remains far from solving the challenge of science research assistance.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Benchmarking AI Agents for Scientific Research Assistance**

A total of **50 papers** were analyzed and organized into a taxonomy with **21 categories**.

Taxonomy Overview

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

- **Benchmark Design and Evaluation Frameworks**
- **AI Agent Architectures and Systems**
- **Human-AI Collaboration and Interaction**
- **Adoption, Usage, and Impact Studies**
- **Research Quality and Evaluation Methodologies**
- **Domain-Specific Applications and Tools**
- **Educational and Training Applications**
- **Foundational Concepts and Theoretical Frameworks**

Complete Taxonomy Tree

- Benchmarking AI Agents for Scientific Research Assistance Survey Taxonomy
- Benchmark Design and Evaluation Frameworks
  - Comprehensive Multi-Task Research Benchmarks ★ (7 papers)
  - [0] AstaBench: Rigorous Benchmarking of AI Agents with a Scientific Research Suite (Anon et al., 2026) View paper
  - [1] Eaira: Establishing a methodology for evaluating ai models as scientific research assistants (CAPPELLO, 2025) View paper
  - [6] Scieval: A multi-level large language model evaluation benchmark for scientific research (Sun, 2024) View paper
  - [11] MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research (Chen Hui, 2025) View paper
  - [12] Mlgym: A new framework and benchmark for advancing ai research agents (Nathani, 2025) View paper
  - [14] Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery (Chen Ziru, 2024) View paper
  - [17] Benchmarking Large Language Models As AI Research Agents (Huang Qian, 2023) View paper
  - Specialized Task Benchmarks (5 papers)
  - [8] Scireplicate-bench: Benchmarking llms in agent-driven algorithmic reproduction from research papers (Yan, 2025) View paper
  - [9] NewtonBench: Benchmarking Generalizable Scientific Law Discovery in LLM Agents (Zheng, 2025) View paper
  - [29] Can AI Validate Science? Benchmarking LLMs for Accurate Scientific Claim â Evidence Reasoning (Cao YuPeng, 2025) View paper
  - [36] EXP-Bench: Can AI Conduct AI Research Experiments? (Liu Jiachen, 2025) View paper
  - [48] PaperArena: An Evaluation Benchmark for Tool-Augmented Agentic Reasoning on Scientific Literature (Wang Daoyu, 2025) View paper
  - Domain-Specific Scientific Benchmarks (3 papers)
  - [7] Benchmarking AI scientists in omics data-driven biological research (Erpai Luo, 2025) View paper
  - [13] BioDSA-1K: Benchmarking Data Science Agents for Biomedical Research (Wang Zifeng, 2025) View paper
  - [22] EarthSE: A Benchmark Evaluating Earth Scientific Exploration Capability for Large Language Models (W Xu, 2025) View paper

- Research Synthesis and Literature Analysis Evaluation (2 papers)
- [19] DeepScholar-Bench: A Live Benchmark and Automated Evaluation for Generative Research Synthesis (Arabzadeh, 2025) View paper
- [27] Towards artificial intelligence research assistant for expert-involved learning (Liu, 2025) View paper
- Reinforcement Learning Research Agent Evaluation (2 papers)
- [33] AI Research Agents for Machine Learning: Search, Exploration, and Generalization in MLE-bench (Hambardzumyan, 2025) View paper
- [44] Re-bench: Evaluating frontier ai r&d capabilities of language model agents against human experts (Wijk, 2024) View paper

- AI Agent Architectures and Systems
  - Autonomous End-to-End Research Systems (5 papers)
  - [2] AI Agents for Deep Scientific Research (R Zhou, 2025) View paper
  - [4] Towards an AI co-scientist (Gottweis, 2025) View paper
  - [10] Evaluating sakana's ai scientist for autonomous research: Wishful thinking or an emerging reality towards' artificial research intelligence'(ari) (Beel, 2025) View paper
  - [18] AI-Researcher: Autonomous Scientific Innovation (Tang, 2025) View paper
  - [31] Evaluating Sakana's AI Scientist: Bold Claims, Mixed Results, and a Promising Future? (Joeran Beel, 2025) View paper
  - Deep Research and Information Retrieval Agents (2 papers)
  - [5] Deepresearcher: Scaling deep research via reinforcement learning in real-world environments (Zheng Yu-Xiang, 2025) View paper
  - [23] Deep research agents: A systematic examination and roadmap (Huang Yuxuan, 2025) View paper
  - Experimental Rigor and Methodological Control Systems (1 papers)
  - [30] Curie: Toward Rigorous and Automated Scientific Experimentation with AI Agents (Liu Jiachen, 2025) View paper
  - Collaborative and Multi-Agent Research Frameworks (1 papers)
  - [39] Agentrxiv: Towards collaborative autonomous research (Schmidgall, 2025) View paper
- Human-AI Collaboration and Interaction
  - Collaborative Intelligence Frameworks (2 papers)
  - [3] Collaborative Intelligence: A scoping review of current applications (Emma Schleiger, 2024) View paper
  - [28] Exploring the role of human-AI collaboration in solving scientific problems (Dazhen Tong, 2025) View paper
  - Mixed-Initiative and Co-Creation Methods (1 papers)
  - [20] Mixed-Initiative Methods for Co-Creation in Scientific Research (Marissa Radensky, 2024) View paper
  - Peer Review and Meta-Review Support (1 papers)
  - [43] Metawriter: Exploring the potential and perils of ai writing support in scientific peer review (Lu Sun, 2024) View paper
- Adoption, Usage, and Impact Studies
  - Adoption Patterns and User Behavior (2 papers)
  - [26] Adoption of AI writing tools among academic researchers: A Theory of Reasoned Action approach (Mohammed A Al-bukhrani, 2025) View paper
  - [37] Evaluating User Interactions and Adoption Patterns of Generative AI in Health Care Occupations Using Claude: Cross-Sectional Study (Gabriel Alain, 2025) View paper
  - Implementation and Deployment Reports (1 papers)
  - [24] The Impact of AI-Driven Chatbot Assistance on Protocol Development and Clinical Research Engagement: An Implementation Report (Kusal Weerasinghe, 2025) View paper
- Research Quality and Evaluation Methodologies (2 papers)
  - [35] Research evaluation with ChatGPT: is it age, country, length, or field biased? (Mike Thelwall, 2025) View paper
  - [47] Research quality evaluation by AI in the era of large language models: advantages, disadvantages, and systemic effectsâ□□An opinion paper (Mike Thelwall, 2025) View paper
- Domain-Specific Applications and Tools
  - Drug Discovery and Biomedical Research Agents (1 papers)
  - [38] AI Agents in Drug Discovery (Srijit Seal, 2025) View paper
  - Clinical Decision Support and Healthcare AI (5 papers)
  - [40] Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges (Qaiser Abbas, 2025) View paper
  - [41] Consensus statements on the current landscape of artificial intelligence applications in endoscopy, addressing roadblocks, and advancing artificial intelligence in â□¦ (AAT Force, 2025) View paper
  - [42] CANAIRI: The collaboration for translational artificial intelligence trials in healthcare (Melissa Mccradden, 2025) View paper
  - [49] International evaluation of an AI system for breast cancer screening (McKinney, 2020) View paper
  - [50] Evaluating artificial intelligence in medicine: phases of clinical research (Yoonyoung Park, 2020) View paper
  - Data Science Automation and Tool Evaluation (1 papers)
  - [45] Measuring Data Science Automation: A Survey of Evaluation Tools for AI Assistants and Agents (Hernandez-Orallo, 2025) View paper
  - Specialized Scientific Computing and Visualization (2 papers)
  - [34] Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE) (B. Kelly, 2022) View paper
  - [46] Software Tools and Evaluation Models for Visual Analytics: Trends, Challenges, and Future Directions (Srikanth Lakumarapu, 2025) View paper
- Educational and Training Applications (3 papers)
  - [16] Design and assessment of AI-based learning tools in higher education: A systematic review (Jihao Luo, 2025) View paper
  - [21] Using leadership to leverage ChatGPT and artificial intelligence for undergraduate and postgraduate research supervision (Joseph Crawford, 2023) View paper
  - [32] CleverCOMSRL: implementation of an AI computer-aided design system in the context of the cognitive science paradigm for the research training process (Olena Hrybiuk, 2024) View paper
- Foundational Concepts and Theoretical Frameworks (2 papers)
  - [15] Integrated Systems for Computational Scientific Discovery (Pat Langley, 2024) View paper
  - [25] Agentic ai for scientific discovery: A survey of progress, challenges, and future directions (Gridach, 2025) View paper

Narrative

Core task: benchmarking AI agents for scientific research assistance. The field has evolved into a rich ecosystem organized around eight major branches. Benchmark Design and Evaluation Frameworks focuses on creating comprehensive multi-task testbeds that assess agents across diverse research activities, from literature review to experimental design, as exemplified by works like AstaBench[0] and SciEval[6]. AI Agent Architectures and Systems explores the underlying technical implementations, including multi-agent collaboration and tool integration strategies seen in systems such as DeepResearcher[5] and AI Co-Scientist[4]. Human-AI Collaboration and Interaction examines how researchers and AI systems work together, while Adoption, Usage, and Impact Studies track real-world deployment patterns. Research Quality and Evaluation Methodologies address the challenge of assessing scientific output validity, and Domain-Specific Applications span areas from drug discovery to materials science. Educational and Training Applications consider how these tools support learning, and Foundational Concepts provide theoretical grounding for agent capabilities and limitations.

A particularly active tension runs between holistic benchmarks that test end-to-end research workflows versus narrower evaluations targeting specific subtasks like literature synthesis or experimental replication. Works such as MLR-Bench[11] and ScienceAgentBench[14] illustrate this spectrum, with some emphasizing breadth across research stages and others drilling into reproducibility or domain expertise. AstaBench[0] sits within the Comprehensive Multi-Task Research Benchmarks cluster, sharing with neighbors like SciEval[6] and MLGym[12] an emphasis on evaluating agents across multiple interconnected research activities rather than isolated skills. Compared to more specialized efforts like SciReplicate-Bench[8], which targets experimental reproducibility, or domain-focused benchmarks such as NewtonBench[9], AstaBench[0] adopts a broader scope that mirrors the multifaceted nature of real scientific inquiry. This positioning reflects an ongoing debate about whether generalist or specialist evaluation paradigms better capture the capabilities needed for meaningful research assistance.

Related Works in Same Category

The following **6 sibling papers** share the same taxonomy leaf node with the original paper:

1. Eaira: Establishing a methodology for evaluating ai models as scientific research assistants

Authors: CAPPELLO, FRANCK, Madireddy, Sandeep, Franck Cappello, et al. (63 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Recent advancements have positioned AI, and particularly Large Language Models (LLMs), as transformative tools for scientific research, capable of addressing complex tasks that require reasoning, problem-solving, and decision-making. Their exceptional capabilities suggest their potential as scientific research assistants but also highlight the need for holistic, rigorous, and domain-specific evaluation to assess effectiveness in real-world scientific applications. This paper describes a multifac...

Relationship Analysis

Both papers belong to the Comprehensive Multi-Task Research Benchmarks category, focusing on evaluating AI agents across end-to-end scientific research workflows. They overlap in assessing literature understanding, code execution, data analysis, and multi-step reasoning capabilities for scientific research assistance. However, AstaBench provides a single integrated suite with 2400+ problems, standardized tools (including production-grade search), and cost-aware evaluation across 11 benchmarks, while EAIRA proposes a broader four-part methodology (MCQ, Open Response, Lab-style, Field-style experiments) emphasizing real-world researcher-LLM interactions and safety considerations across multiple evaluation paradigms rather than a unified benchmark suite.

---

2. Scieval: A multi-level large language model evaluation benchmark for scientific research

Authors: Sun, Liangtai, Han, Yang, Liangtai Sun, et al. (17 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Recently, there has been growing interest in using Large Language Models (LLMs) for scientific research. Numerous benchmarks have been proposed to evaluate the ability of LLMs for scientific research. However, current benchmarks are mostly based on pre-collected objective questions. This design suffers from data leakage problem and lacks the evaluation of subjective Q/A ability. In this paper, we propose SciEval, a comprehensive and multi-disciplinary evaluation benchmark to address these issues...

Relationship Analysis

Both papers belong to the Comprehensive Multi-Task Research Benchmarks category, evaluating AI agents across multiple scientific research tasks. They overlap in assessing literature understanding, scientific reasoning, and code execution capabilities within scientific domains. However, AstaBench focuses on end-to-end research workflows with production-grade search tools and real user-derived tasks, while SciEval emphasizes multi-level cognitive evaluation (based on Bloom's taxonomy) with dynamic data generation to prevent leakage and includes experimental design assessment.

---

3. MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research

Authors: Chen Hui, Xiong Miao, Hui Chen, Lu Yujie, Miao Xiong, et al. (22 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Recent advancements in AI agents have demonstrated their growing potential to drive and support scientific discovery. In this work, we introduce MLR-Bench, a comprehensive benchmark for evaluating AI agents on open-ended machine learning research. MLR-Bench includes three key components: (1) 201 research tasks sourced from NeurIPS, ICLR, and ICML workshops covering diverse ML topics; (2) MLR-Judge, an automated evaluation framework combining LLM-based reviewers with carefully designed review rub...

Relationship Analysis

Both papers belong to the Comprehensive Multi-Task Research Benchmarks category, evaluating AI agents across end-to-end scientific research workflows. They overlap in assessing agents' capabilities for literature review, experimentation, code execution, and report generation within scientific domains. However, AstaBench focuses on 2400+ problems spanning the full discovery process with production-grade search tools and cost-aware evaluation across multiple scientific domains, while MLR-Bench specifically targets 201 open-ended machine learning research tasks from ML conferences with a modular four-stage agent scaffold and LLM-based review evaluation.

---

4. Mlgym: A new framework and benchmark for advancing ai research agents

Authors: Nathani, Deepak, Madaan, Lovish, Deepak Nathani, et al. (43 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

We introduce Meta MLGym and MLGym-Bench, a new framework and benchmark for evaluating and developing LLM agents on AI research tasks. This is the first Gym environment for machine learning (ML) tasks, enabling research on reinforcement learning (RL) algorithms for training such agents. MLGym-bench consists of 13 diverse and open-ended AI research tasks from diverse domains such as computer vision, natural language processing, reinforcement learning, and game theory. Solving these tasks requires ...

Relationship Analysis

Both papers belong to the Comprehensive Multi-Task Research Benchmarks category, evaluating AI agents across end-to-end scientific research workflows. They overlap in assessing agents on literature understanding, code execution, data analysis, and multi-step research tasks using standardized environments and cost-aware evaluation. However, AstaBench focuses on 2400+ problems spanning the full scientific discovery process with production-grade literature search tools and real user-derived tasks, while MLGym provides a Gym-based RL framework for 13 open-ended ML research tasks emphasizing algorithmic reasoning (RL, game theory, SAT) and flexible artifact evaluation (model weights, algorithms, strategies).

---

5. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery

Authors: Chen Ziru, Chen Shijie, Ziru Chen, Ning, Yuting, et al. (43 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

The advancements of large language models (LLMs) have piqued growing interest in developing LLM-based language agents to automate scientific discovery end-to-end, which has sparked both excitement and skepticism about their true capabilities. In this work, we call for rigorous assessment of agents on individual tasks in a scientific workflow before making bold claims on end-to-end automation. To this end, we present ScienceAgentBench, a new benchmark for evaluating language agents for data-drive...

Relationship Analysis

Both papers belong to the Comprehensive Multi-Task Research Benchmarks category, evaluating AI agents across end-to-end scientific research workflows. They overlap in assessing agents on literature understanding, code generation, data analysis, and providing standardized evaluation frameworks with controlled tools. However, AstaBench emphasizes production-grade search tools with 2400+ problems spanning the full discovery process and cost-aware leaderboards, while ScienceAgentBench focuses on 102 tasks extracted from peer-reviewed publications with rigorous expert validation and data contamination mitigation strategies.

---

6. Benchmarking Large Language Models As AI Research Agents

Authors: Huang Qian, Qian Huang, Vora, Jian, Jian Vora, et al. (11 authors total) | **Year/Venue**: 2023 • arXiv.org | **URL**: View paper

Abstract

N/A

Relationship Analysis

Both papers belong to the comprehensive multi-task research benchmarks category, evaluating AI agents across end-to-end scientific research workflows. They overlap in testing agents on literature understanding, code execution, and data analysis tasks within scientific domains. However, the original paper (AstaBench) provides production-grade search tools with date-restricted corpus access and emphasizes cost-aware evaluation across 2400+ problems spanning the full discovery process, while the candidate paper (MLAgentBench) focuses specifically on machine learning experimentation tasks (13 tasks) with emphasis on iterative model improvement and does not provide standardized literature search tools or systematic cost accounting.

Contributions Analysis

Overall novelty summary. The paper proposes AstaBench, a comprehensive benchmark suite for evaluating AI agents across the full scientific research workflow, comprising 2400+ problems. It resides in the 'Comprehensive Multi-Task Research Benchmarks' leaf alongside six sibling papers, including SciEval, MLR-Bench, and ScienceAgentBench. This leaf represents a moderately populated research direction within a taxonomy of 50 papers across 21 leaf nodes, indicating active but not overcrowded interest in holistic, multi-task evaluation frameworks that assess end-to-end research capabilities rather than isolated subtasks.

The taxonomy reveals neighboring leaves focused on 'Specialized Task Benchmarks' (targeting specific subtasks like code reproduction or hypothesis validation) and 'Domain-Specific Scientific Benchmarks' (evaluating agents within particular scientific domains). AstaBench's positioning emphasizes breadth across research stages—literature review, experimental design, data analysis—distinguishing it from narrower efforts like SciReplicate-Bench (experimental reproducibility) or domain-focused benchmarks such as NewtonBench. The taxonomy structure shows an ongoing tension between generalist multi-task evaluations and specialist assessments, with AstaBench aligning with the former approach.

Among 29 candidates examined, the 'AstaBench benchmark suite' contribution shows one refutable candidate out of nine examined, suggesting some prior work overlap in comprehensive research benchmarking. The 'Asta Environment with production-grade search tools' contribution examined 10 candidates with none clearly refuting it, indicating relative novelty in providing standardized, reproducible agent tooling. The 'agent-eval Toolkit and comprehensive agents suite' contribution similarly examined 10 candidates with no clear refutations, suggesting this infrastructure component addresses a less-explored gap in baseline agent provision and rapid prototyping interfaces.

Based on this limited search scope of 29 semantically similar papers, the work appears to offer incremental advances in benchmark comprehensiveness and tooling standardization within an active research area. The analysis covers top-K semantic matches and does not constitute an exhaustive literature review, leaving open the possibility of additional relevant prior work in adjacent communities or recent preprints not captured by the search strategy.

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

Contribution 1: AstaBench benchmark suite for scientific research agents

Description: The authors introduce AstaBench, a comprehensive benchmark suite designed to holistically evaluate AI agents' capabilities in scientific research. It includes over 2400 problems covering the full research pipeline across multiple domains, with many tasks inspired by real user requests from deployed Asta agents.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery

URL: View paper

Prior Art Analysis

ScienceAgentBench[14] demonstrates that a similar benchmark for evaluating AI agents in scientific research already exists. Both benchmarks extract tasks from peer-reviewed publications, engage subject matter experts for validation, evaluate agents on data-driven scientific discovery tasks, and provide comprehensive evaluation frameworks. ScienceAgentBench[14] was published at ICLR 2025 and presents 102 tasks from 44 publications across four scientific disciplines, with rigorous evaluation metrics and expert validation—features that overlap substantially with the original paper's claimed novelty.

Evidence

Evidence 1 - **Rationale**: Both benchmarks aim to holistically evaluate agents across the scientific research workflow. ScienceAgentBench[14] evaluates essential tasks in data-driven discovery including model development, data analysis, and visualization, which overlaps with the original paper's claim of covering the full scientific discovery process. - **Original**: guided by our principles, we present astabench1 (section 3), a more rigorous agent benchmark suite that isa holistic measure of scientific research, which exercises a broad spectrum of skillsincluding literature understanding, data understanding, planning, tool use, coding, and search-and comprises ... - **Candidate**: in this work, we contend that for a language agent to fully automate data-driven discovery, it must be able to complete all essential tasks in the workflow, such as model development, data analysis, and visualization. thus, we advocate careful evaluations of the agents' performance on these tasks, b...

---

2. Survey on evaluation of llm-based agents

URL: View paper

Brief Assessment

LLM-Based Agents Survey[52] provides a broad survey of evaluation methodologies for LLM-based agents across multiple domains, including scientific agents as one category among many. While it discusses scientific agent benchmarks, it does not claim to introduce a comprehensive benchmark suite specifically designed for holistic evaluation of the full scientific research pipeline like AstaBench does.

---

3. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents

URL: View paper

Brief Assessment

DiscoveryWorld[53] focuses on end-to-end scientific discovery in a virtual environment with simulated experiments, while AstaBench evaluates agents across the full research pipeline using production-grade tools and real scientific literature. The candidate addresses virtual simulation environments rather than comprehensive benchmarking of research assistance capabilities.

---

4. Towards an AI co-scientist: A multi-agent system for scientific discovery

URL: View paper

Brief Assessment

Multi-Agent Co-Scientist[51] focuses on a multi-agent system for hypothesis generation and scientific discovery through debate and evolution mechanisms, not on creating benchmarks for evaluating AI agents across the research pipeline.

---

5. Towards an AI co-scientist

URL: View paper

Brief Assessment

AI Co-Scientist[4] focuses on a multi-agent system for hypothesis generation and scientific discovery, not on creating a benchmark suite for evaluating AI agents in scientific research. The systems serve different purposes in the scientific research pipeline.

---

6. Mlagentbench: Evaluating language agents on machine learning experimentation

URL: View paper

Brief Assessment

MLAgentBench[54] focuses specifically on machine learning experimentation tasks (e.g., improving model performance on CIFAR-10), not the broader scientific research pipeline that AstaBench covers. The candidate addresses a narrower domain of ML experimentation rather than holistic scientific research assistance.

---

7. Auto-Bench: An Automated Benchmark for Scientific Discovery in LLMs

URL: View paper

Brief Assessment

Auto-Bench[56] focuses on causal graph discovery for scientific discovery evaluation, while AstaBench provides a comprehensive benchmark suite covering the full research pipeline including literature understanding, code execution, data analysis, and end-to-end discovery across multiple domains.

---

8. AI Agents for Deep Scientific Research

URL: View paper

Brief Assessment

AI Agents Deep Research[2] surveys existing AI research agents and their evaluation methods but does not present a comprehensive benchmark suite for holistically evaluating scientific research agents across the full research pipeline. The candidate focuses on reviewing architectures and capabilities rather than introducing a competing benchmark.

---

9. Mlgym: A new framework and benchmark for advancing ai research agents

URL: View paper

Brief Assessment

MLGym[12] focuses on AI research tasks in machine learning domains (computer vision, NLP, RL, game theory) rather than the broader scientific research pipeline that AstaBench covers. MLGym[12] does not include literature understanding, search, or comprehensive research workflow tasks that are central to AstaBench.

Contribution 2: Asta Environment with production-grade search tools

Description: The authors develop the Asta Environment, which provides the first realistic and reproducible scientific research environment for agents. It features production-grade search tools with date-restricted access to scientific literature, enabling controlled comparison of agents while accounting for confounding variables.

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. Evaluating Retrieval-Augmented Generation Agents for Autonomous Scientific Discovery in Astrophysics

URL: View paper

Brief Assessment

RAG Agents Astrophysics[58] focuses on retrieval-augmented generation systems for cosmology research, not on creating general-purpose agent environments with production-grade search tools for controlled evaluation across diverse scientific domains.

---

2. Kwaiagents: Generalized information-seeking agent system with large language models

URL: View paper

Brief Assessment

KwaiAgents[66] focuses on a generalized information-seeking agent system with search-browse toolkits for web navigation and entity linking, not on providing a reproducible scientific research environment with date-restricted access to scientific literature as described in the original contribution.

---

3. Mcp-bench: Benchmarking tool-using llm agents with complex real-world tasks via mcp servers

URL: View paper

Brief Assessment

MCP-Bench[61] focuses on tool-using LLM agents across diverse domains (finance, travel, scientific computing) via MCP servers, not specifically on scientific research environments with production-grade literature search tools like the original paper's Asta Environment.

---

4. Nanostructured material design via a retrieval-augmented generation (rag) approach: Bridging laboratory practice and scientific literature

URL: View paper

Brief Assessment

RAG Nanostructured Materials[60] focuses on a domain-specific RAG system for nanostructured materials literature, not on creating a general scientific research environment with production-grade search tools for controlled agent evaluation.

---

5. PaSa: An LLM Agent for Comprehensive Academic Paper Search

URL: View paper

Brief Assessment

PaSa[65] focuses on building an LLM agent for academic paper search using Google and arXiv APIs, not on creating a reproducible scientific research environment with production-grade search tools and date-restricted access for controlled agent evaluation.

---

6. EvoPat: A Multi-LLM-based Patents Summarization and Analysis Agent

URL: View paper

Brief Assessment

EvoPat[62] focuses on patent analysis using retrieval-augmented generation with local databases of patents and literature, not on providing a reproducible scientific research environment with production-grade search tools for controlled agent evaluation.

---

7. TourSynbio-Search: A Large Language Model Driven Agent Framework for Unified Search Method for Protein Engineering

URL: View paper

Brief Assessment

TourSynbio-Search[64] focuses on protein engineering databases (UniProt, PDB) and biological literature search, not on general scientific research environments with controlled, reproducible evaluation tools for AI agents.

---

8. Spar: Scholar paper retrieval with llm-based agents for enhanced academic search

URL: View paper

Brief Assessment

SPAR[59] focuses on a multi-agent retrieval framework for academic search, not on providing a standardized evaluation environment with production-grade tools for agent benchmarking. The systems serve different purposes: SPAR is a retrieval system, while Asta Environment is an evaluation infrastructure.

---

9. Open-Source Agentic Hybrid RAG Framework for Scientific Literature Review

URL: View paper

Brief Assessment

Agentic Hybrid RAG[63] focuses on a hybrid retrieval system combining graph and vector search for literature review, not on providing a standardized agent environment with production-grade search tools for controlled agent evaluation and benchmarking.

---

10. LITERAS: Biomedical literature review and citation retrieval agents

URL: View paper

Brief Assessment

LITERAS[57] focuses on biomedical literature review and citation retrieval. The provided context is too limited (only fragments) to assess whether it offers production-grade search tools comparable to the Asta Environment's controlled, date-restricted scientific corpus access.

Contribution 3: agent-eval Toolkit and comprehensive agents suite

Description: The authors present the agent-eval toolkit for standardized agent evaluation with time-invariant cost tracking, alongside the agent-baselines suite containing nine science-optimized Asta agent classes and numerous baselines. This represents the most comprehensive standardized agents suite for scientific research tasks.

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. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery

URL: View paper

Brief Assessment

ScienceAgentBench[14] does not present a standardized evaluation toolkit with time-invariant cost tracking or a comprehensive suite of agent classes. While it evaluates multiple agents, it does not claim to provide a reusable toolkit or agent baselines suite as the original paper does.

---

2. Evaluating LLM Agent Adherence to Hierarchical Safety Principles: A Lightweight Benchmark for Probing Foundational Controllability Components

URL: View paper

Brief Assessment

LLM Agent Safety Principles[69] focuses on evaluating agent adherence to safety principles through a lightweight benchmark, not on providing standardized agent evaluation toolkits with time-invariant cost tracking or comprehensive agent suites for scientific research tasks.

3. Benchmarking and management accounting: A framework for research.

URL: View paper

Brief Assessment

Benchmarking Management Accounting[74] focuses on benchmarking and management accounting frameworks in business contexts, not AI agent evaluation toolkits or standardized agent suites for scientific research tasks.

4. Multi-Agent Penetration Testing AI for the Web

URL: View paper

Brief Assessment

Multi-Agent Penetration Testing[67] focuses on web application security assessment using multi-agent systems for vulnerability detection, not on standardized agent evaluation toolkits with time-invariant cost tracking or comprehensive agent baselines for scientific research tasks.

5. MDPs with a State Sensing Cost

URL: View paper

Brief Assessment

The candidate paper focuses on MDPs with state sensing costs in sequential decision-making, not on agent evaluation toolkits or standardized benchmarking frameworks for AI agents.

6. Optimizing Control of Wastewater Treatment Plant With Reinforcement Learning: Technical Evaluation of Twin-Delayed Deep Deterministic Policy Gradient Agent

URL: View paper

Brief Assessment

This candidate focuses on reinforcement learning for wastewater treatment plant control using the Twin-Delayed Deep Deterministic Policy Gradient algorithm. It does not address agent evaluation toolkits, standardized benchmarking frameworks, or comprehensive agent suites for scientific research tasks.

7. 360REA: Towards A Reusable Experience Accumulation with 360{\deg} Assessment for Multi-Agent System

URL: View paper

Brief Assessment

360REA[71] focuses on multi-agent performance assessment and experience accumulation mechanisms, not on standardized evaluation toolkits with time-invariant cost tracking or comprehensive agent baselines for benchmarking.

8. Surfer-H Meets Holo1: Cost-Efficient Web Agent Powered by Open Weights

URL: View paper

Brief Assessment

Surfer-H Holo1[68] focuses on web navigation agents and vision-language models for web tasks, not on standardized agent evaluation toolkits with time-invariant cost tracking or comprehensive agent baselines for scientific research tasks.

9. Autonomous Evaluation and Refinement of Digital Agents

URL: View paper

Brief Assessment

Autonomous Agent Refinement[73] focuses on autonomous evaluation and refinement of digital agents through model-based evaluators for web navigation and device control, not on providing a standardized agent evaluation toolkit with time-invariant cost tracking or a comprehensive suite of science-optimized agent classes for research tasks.

10. CostNav: A Navigation Benchmark for Cost-Aware Evaluation of Embodied Agents

URL: View paper

Brief Assessment

CostNav[72] focuses on economic evaluation of navigation agents for autonomous delivery robots, not on general agent evaluation toolkits with time-invariant cost tracking or standardized benchmarking frameworks for scientific research tasks.

Appendix: Text Similarity Detection

Textual similarity detection checked 33 papers and found 2 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. Benchmarking Large Language Models As AI Research Agents

Detected in: Core Task (sibling)

⚠ **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] AstaBench: Rigorous Benchmarking of AI Agents with a Scientific Research Suite View paper
- [1] Eaira: Establishing a methodology for evaluating ai models as scientific research assistants View paper
- [2] AI Agents for Deep Scientific Research View paper
- [3] Collaborative Intelligence: A scoping review of current applications View paper
- [4] Towards an AI co-scientist View paper
- [5] Deepresearcher: Scaling deep research via reinforcement learning in real-world environments View paper
- [6] Scieval: A multi-level large language model evaluation benchmark for scientific research View paper
- [7] Benchmarking AI scientists in omics data-driven biological research View paper
- [8] Scireplicate-bench: Benchmarking llms in agent-driven algorithmic reproduction from research papers View paper
- [9] NewtonBench: Benchmarking Generalizable Scientific Law Discovery in LLM Agents View paper

- [10] Evaluating sakana's ai scientist for autonomous research: Wishful thinking or an emerging reality towards' artificial research intelligence'(ari) View paper
- [11] MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research View paper
- [12] Mlgym: A new framework and benchmark for advancing ai research agents View paper
- [13] BioDSA-1K: Benchmarking Data Science Agents for Biomedical Research View paper
- [14] Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery View paper
- [15] Integrated Systems for Computational Scientific Discovery View paper
- [16] Design and assessment of AI-based learning tools in higher education: A systematic review View paper
- [17] Benchmarking Large Language Models As AI Research Agents View paper
- [18] AI-Researcher: Autonomous Scientific Innovation View paper
- [19] DeepScholar-Bench: A Live Benchmark and Automated Evaluation for Generative Research Synthesis View paper
- [20] Mixed-Initiative Methods for Co-Creation in Scientific Research View paper
- [21] Using leadership to leverage ChatGPT and artificial intelligence for undergraduate and postgraduate research supervision View paper
- [22] EarthSE: A Benchmark Evaluating Earth Scientific Exploration Capability for Large Language Models View paper
- [23] Deep research agents: A systematic examination and roadmap View paper
- [24] The Impact of AI-Driven Chatbot Assistance on Protocol Development and Clinical Research Engagement: An Implementation Report View paper
- [25] Agentic ai for scientific discovery: A survey of progress, challenges, and future directions View paper
- [26] Adoption of AI writing tools among academic researchers: A Theory of Reasoned Action approach View paper
- [27] Towards artificial intelligence research assistant for expert-involved learning View paper
- [28] Exploring the role of human-AI collaboration in solving scientific problems View paper
- [29] Can AI Validate Science? Benchmarking LLMs for Accurate Scientific Claim â□□ Evidence Reasoning View paper
- [30] Curie: Toward Rigorous and Automated Scientific Experimentation with AI Agents View paper
- [31] Evaluating Sakana's AI Scientist: Bold Claims, Mixed Results, and a Promising Future? View paper
- [32] CleverCOMSRL: implementation of an AI computer-aided design system in the context of the cognitive science paradigm for the research training process View paper
- [33] AI Research Agents for Machine Learning: Search, Exploration, and Generalization in MLE-bench View paper
- [34] Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE) View paper
- [35] Research evaluation with ChatGPT: is it age, country, length, or field biased? View paper
- [36] EXP-Bench: Can AI Conduct AI Research Experiments? View paper
- [37] Evaluating User Interactions and Adoption Patterns of Generative AI in Health Care Occupations Using Claude: Cross-Sectional Study View paper
- [38] AI Agents in Drug Discovery View paper
- [39] Agentrxiv: Towards collaborative autonomous research View paper
- [40] Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges View paper
- [41] Consensus statements on the current landscape of artificial intelligence applications in endoscopy, addressing roadblocks, and advancing artificial intelligence in â□¦ View paper
- [42] CANAIRI: The collaboration for translational artificial intelligence trials in healthcare View paper
- [43] Metawriter: Exploring the potential and perils of ai writing support in scientific peer review View paper
- [44] Re-bench: Evaluating frontier ai r&d capabilities of language model agents against human experts View paper
- [45] Measuring Data Science Automation: A Survey of Evaluation Tools for AI Assistants and Agents View paper
- [46] Software Tools and Evaluation Models for Visual Analytics: Trends, Challenges, and Future Directions View paper
- [47] Research quality evaluation by AI in the era of large language models: advantages, disadvantages, and systemic effectsâ□□An opinion paper View paper
- [48] PaperArena: An Evaluation Benchmark for Tool-Augmented Agentic Reasoning on Scientific Literature View paper
- [49] International evaluation of an AI system for breast cancer screening View paper
- [50] Evaluating artificial intelligence in medicine: phases of clinical research View paper
- [51] Towards an AI co-scientist: A multi-agent system for scientific discovery View paper
- [52] Survey on evaluation of llm-based agents View paper
- [53] Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents View paper
- [54] Mlagentbench: Evaluating language agents on machine learning experimentation View paper
- [55] From ai for science to agentic science: A survey on autonomous scientific discovery View paper
- [56] Auto-Bench: An Automated Benchmark for Scientific Discovery in LLMs View paper
- [57] LITERAS: Biomedical literature review and citation retrieval agents View paper
- [58] Evaluating Retrieval-Augmented Generation Agents for Autonomous Scientific Discovery in Astrophysics View paper
- [59] Spar: Scholar paper retrieval with llm-based agents for enhanced academic search View paper
- [60] Nanostructured material design via a retrieval-augmented generation (rag) approach: Bridging laboratory practice and scientific literature View paper
- [61] Mcp-bench: Benchmarking tool-using llm agents with complex real-world tasks via mcp servers View paper
- [62] EvoPat: A Multi-LLM-based Patents Summarization and Analysis Agent View paper
- [63] Open-Source Agentic Hybrid RAG Framework for Scientific Literature Review View paper
- [64] TourSynbio-Search: A Large Language Model Driven Agent Framework for Unified Search Method for Protein Engineering View paper
- [65] PaSa: An LLM Agent for Comprehensive Academic Paper Search View paper
- [66] Kwaiagents: Generalized information-seeking agent system with large language models View paper
- [67] Multi-Agent Penetration Testing AI for the Web View paper
- [68] Surfer-H Meets Holo1: Cost-Efficient Web Agent Powered by Open Weights View paper
- [69] Evaluating LLM Agent Adherence to Hierarchical Safety Principles: A Lightweight Benchmark for Probing Foundational Controllability Components View paper
- [70] Optimizing Control of Wastewater Treatment Plant With Reinforcement Learning: Technical Evaluation of Twin-Delayed Deep Deterministic Policy Gradient Agent View paper

- [71] 360REA: Towards A Reusable Experience Accumulation with 360{\deg} Assessment for Multi-Agent System View paper
- [72] CostNav: A Navigation Benchmark for Cost-Aware Evaluation of Embodied Agents View paper
- [73] Autonomous Evaluation and Refinement of Digital Agents View paper
- [74] Benchmarking and management accounting: A framework for research. View paper
- [75] MDPs with a State Sensing Cost View paper