SCIENCEAGENTBENCH: TOWARD RIGOROUS ASSESSMENT OF LANGUAGE AGENTS FOR DATA-DRIVEN SCIENTIFIC DISCOVERY

Ziru Chen[1*], **Shijie Chen**[1*], **Yuting Ning**[1], **Qianheng Zhang**[3], **Boshi Wang**[1], **Botao Yu**[1], **Yifei Li**[1], **Zeyi Liao**[1], **Chen Wei**[3], **Zitong Lu**[4], **Vishal Dey**[1], **Mingyi Xue**[5], **Frazier N. Baker**[1,6], **Benjamin Burns**[1], **Daniel Adu-Ampratwum**[2], **Xuhui Huang**[5], **Xia Ning**[1,2,6], **Song Gao**[3], **Yu Su**[1], **Huan Sun**[1*]
[1]Department of Computer Science and Engineering, OSU [2]College of Pharmacy, OSU
[3]Department of Geography, UW–Madison [4]Department of Psychology, OSU
[5]Department of Chemistry, UW–Madison [6]Department of Biomedical Informatics, OSU
*Correspondence to: {chen.8336,chen.10216,sun.397}@osu.edu
Website: https://osu-nlp-group.github.io/ScienceAgentBench/

ABSTRACT

The advancements of language 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 the true capabilities of such agents. In this work, we argue that for an agent to fully automate scientific discovery, it must be able to complete all essential tasks in the workflow. Thus, 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-driven scientific discovery. To ensure the scientific authenticity and real-world relevance of our benchmark, we extract 102 tasks from 44 peer-reviewed publications in four disciplines and engage nine subject matter experts to validate them. We unify the target output for every task to a self-contained Python program file and employ an array of evaluation metrics to examine the generated programs, execution results, and costs. Each task goes through multiple rounds of manual validation by annotators and subject matter experts to ensure its annotation quality and scientific plausibility. We also propose two effective strategies to mitigate data contamination concerns. Using our benchmark, we evaluate five open-weight and proprietary LLMs, each with three frameworks: direct prompting, OpenHands CodeAct, and self-debug. Given three attempts for each task, the best-performing agent can only solve 32.4% of the tasks independently and 34.3% with expert-provided knowledge. These results underscore the limited capacities of current language agents in generating code for data-driven discovery, let alone end-to-end automation for scientific research.

1 INTRODUCTION

Large language models (LLMs) have shown remarkable capabilities beyond text generation, including reasoning (Wei et al., 2022; Yao et al., 2023), tool learning (Schick et al., 2023; Wang et al., 2024a), and code generation (Chen et al., 2021; Yang et al., 2024a). These abilities have piqued significant research interests in developing LLM-based language agents to automate scientific discovery end-to-end. For instance, Majumder et al. (2024a) urge the community to build automated systems for end-to-end *data-driven discovery*, an increasingly important workflow in many disciplines (Hey et al., 2009) that leverages existing datasets to derive new findings. More recently, Lu et al. (2024) claim to have built The AI Scientist, an agent that is capable of automating the entire research workflow, from generating ideas to running experiments and writing papers. This ambitious claim has sparked both excitement and skepticism about the true capabilities of such agents.

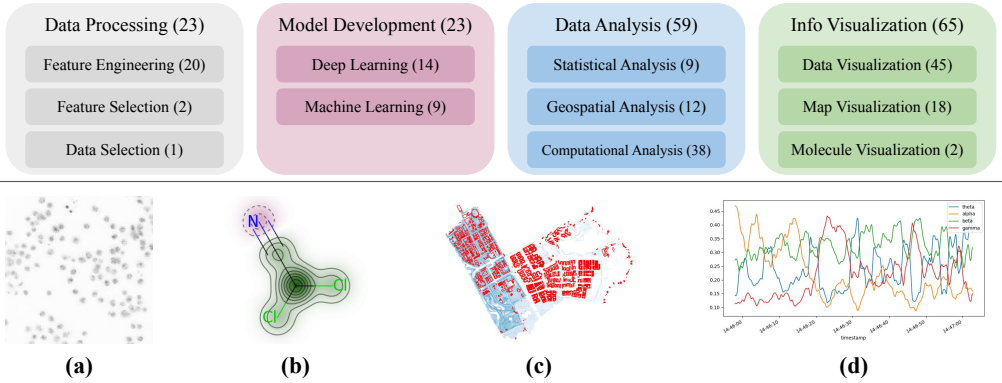

Figure 1: **Top:** Distribution of sub-tasks in ScienceAgentBench. Each task in our benchmark consists of one or more of these sub-tasks and requires successful completion of all sub-tasks to achieve the task goal. **Bottom:** Heterogeneous datasets involved: **(a)** a cell image in Bioinformatics, **(b)** a molecular activity visualization in Computational Chemistry, **(c)** a flooding risk map in Geographical Information Science, and **(d)** an EEG time series in Psychology and Cognitive Neuroscience.

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, before claiming they can automate data-driven discovery end-to-end. Such an assessment strategy helps grasp a more solid understanding of an agent's strengths and limitations than purely relying on end-to-end evaluations, e.g., using an LLM-based reviewer to assess generated papers (Lu et al., 2024). Yet, high-quality benchmarks focusing on individual tasks in real-world scientific workflows are lacking for objective assessment and continued development of agents for data-driven discovery.

To this end, we present ScienceAgentBench, a new benchmark for evaluating language agents for data-driven discovery. The construction of ScienceAgentBench follows three key design principles. **(1) Scientific authenticity through co-design with subject matter experts**: We ensure the authenticity of tasks in our benchmark by directly extracting them from peer-reviewed publications and engaging nine subject matter experts (incl. senior Ph.D. students and professors) from the respective disciplines to validate them. This approach also minimizes the generalization gap for agents developed on our benchmark to real-world scenarios. In total, we curate 102 diverse tasks from 44 peer-reviewed publications in four disciplines: Bioinformatics, Computational Chemistry, Geographical Information Science, and Psychology & Cognitive Neuroscience (Figure 1). **(2) Rigorous graded evaluation**: Reliable evaluation for language agents is notably difficult due to the open-endedness and complexity of data-driven discovery tasks. We first unify the target output for every task as a self-contained Python program, and then employ an array of evaluation metrics that examine the generated programs, execution results (e.g., rendered figures or test set predictions), and costs. We also provide step-by-step rubrics specific to each task to enable graded evaluation. **(3) Careful multi-stage quality control**: Each task goes through multiple rounds of manual validation by annotators and subject matter experts to ensure its quality and scientific plausibility. We also propose two effective strategies to mitigate data contamination concerns due to LLM pre-training.

We comprehensively evaluate five open-weight and proprietary LLMs, each with three frameworks: direct prompting, OpenHands CodeAct (Wang et al., 2024c), and self-debug. Surprisingly, without expert-provided knowledge, Claude-3.5-Sonnet using self-debug can successfully solve **10.8%** more tasks than using OpenHands CodeAct while costing **17** times less API fees. This result resonates with recent findings that agent designs should jointly consider costs and performance to maximize their practical utility (Kapoor et al., 2024). Still, given three attempts for each task, the best agent can only solve 32.4% of the tasks independently and 34.3% of them with expert-provided knowledge. These results also suggest language agents cannot yet automate essential tasks in data-driven discovery nor the research pipelines end-to-end, in contrast to claims in recent work such as Lu et al. (2024).

Despite their current mediocre performance, we believe language agents hold significant potential in augmenting human scientists' productivity: For each task in our benchmark, it takes a trained