

18 August 2026

Measuring AI Scientists: From Exams to Discovery

Yuanqi Du¹, Steven Dillmann², Jon Laurent³, Peter Jansen^{4,5}, Haojun Jia⁶, Ludwig Schmidt², Andrew White³, Kristin Persson⁷, Chenru Duan⁶

1. Microsoft Research New England

2. Stanford University

3. Edison Scientific

4. Allen Institute for AI

5. University of Arizona

6. Deep Principle

7. University of California at Berkeley

Abstract

Large language models and agentic systems are increasingly embedded across the scientific workflow, from literature synthesis and hypothesis generation to code execution, data analysis and writing. This broadening of use exposes a mismatch in evaluation: current benchmarks can test whether models answer difficult scientific questions, but not whether they can advance science itself. Discovery depends on sequences of decisions under uncertainty, in which hypotheses must be made testable, interventions executed reliably, and imperfect evidence interpreted with calibrated restraint. In this Perspective, we argue that AI scientists should be evaluated through discovery episodes rather than isolated problems. A discovery episode records the evolving scientific state, the actions taken, the observations produced, the revisions made and the provenance needed to reproduce and audit the process. This framing separates three coupled capabilities—hypothesis generation, computational or experimental execution, and result interpretation—and motivates evaluations that score trajectories as well as outcomes, preserve failures and null results, and distinguish genuine discovery from rediscovery. Such evaluations are central for developing AI scientists that contribute responsibly to trustworthy scientific discovery.

Measuring AI Scientists: From Exams to Discovery

Yuanqi Du^{1,*}, Steven Dillmann², Jon Laurent³, Peter Jansen^{4,5}, Haojun Jia⁶, Ludwig Schmidt², Andrew White³,
Kristin Persson⁷, and Chenru Duan^{6,*}

¹Microsoft Research New England

²Stanford University

³Edison Scientific

⁴Allen Institute for AI

⁵University of Arizona

⁶Deep Principle

⁷University of California at Berkeley

*Correspondence to: yuanqidu@microsoft.com, duanchenru@gmail.com

Abstract

Large language models and agentic systems are increasingly embedded across the scientific workflow, from literature synthesis and hypothesis generation to code execution, data analysis and writing. This broadening of use exposes a mismatch in evaluation: current benchmarks can test whether models answer difficult scientific questions, but not whether they can advance science itself. Discovery depends on sequences of decisions under uncertainty, in which hypotheses must be made testable, interventions executed reliably, and imperfect evidence interpreted with calibrated restraint. In this Perspective, we argue that AI scientists should be evaluated through discovery episodes rather than isolated problems. A discovery episode records the evolving scientific state, the actions taken, the observations produced, the revisions made and the provenance needed to reproduce and audit the process. This framing separates three coupled capabilities—hypothesis generation, computational or experimental execution, and result interpretation—and motivates evaluations that score trajectories as well as outcomes, preserve failures and null results, and distinguish genuine discovery from rediscovery. Such evaluations are central for developing AI scientists that contribute responsibly to trustworthy scientific discovery.

Introduction

Artificial intelligence is rapidly becoming part of the everyday machinery of scientific research [1]. Large language models (LLMs) and agents now assist with literature synthesis [2], code generation [3], hypothesis generation [4–6], data analysis [7, 8], and, increasingly, interaction with computational and experimental tools [9–14]. This progress extends a longer history of AI for science, from protein structure prediction [15, 16], mathematical discovery [17], and molecular design [18, 19] to autonomous materials discovery [20]. LLMs and agents, however, have changed the scope of AI in science through their accessibility and breadth of adoption: their domain-general interfaces produce fluent, seemingly reasonable responses to almost any scientific request, making their apparent competence far broader and harder to delimit than that of earlier specialized models [21]. The question is therefore no longer only whether specialized models can solve difficult scientific problems, but whether general-purpose AI scientists can participate in the open-ended process of discovery by which new scientific knowledge is produced—and, if they can, how we would know.

Current evaluation practice has not yet caught up with this shift. Widely used science benchmarks measure closed-ended expert knowledge, including multiple-choice questions, short-answer problems, and reasoning traces with textbook-style rubrics. These benchmarks are valuable because they are standardized, scalable, and comparable across models [22–24]. However, as frontier LLMs have begun to saturate many once-difficult benchmarks, the field has moved toward ever-harder expert exams such as Humanity’s Last Exam and FrontierScience [25, 26]. These benchmarks create headroom for new models and expose important gaps in model knowledge and reasoning, but they also sharpen a deeper limitation: answering difficult questions is not the same as doing science. Scientific discovery is an iterative, evidence-producing workflow in which a scientist identifies a meaningful problem, generates hypotheses, translates them into executable experiments or simulations, interprets uncertain and often conflicting results, and decides which directions to pursue next [27]. Each step changes the state of knowledge and informs the next action, making discovery a sequential decision-making process rather than a single act of answering [28]. This loop cannot be fully captured by a static question-answer format, even when the questions are extremely hard. A model may know a physical law, recognize a biological mechanism, or complete an analytical derivation, while still failing to design a robust experiment, recover from tool errors, calibrate uncertainty, detect artifacts, or revise a hypothesis in light of unexpected and possibly noisy evidence. The relevant distinction is therefore not simply between easy and hard benchmarks, but between knowledge and discovery evaluation.

Recent work in AI for science makes this distinction urgent. LLM-based agents have begun to couple hypothesis generation, literature search, code execution, physical experiment control, and result interpretation within longer scientific workflows across chemistry, materials science, and life sciences [6, 10, 12, 20, 29]. These demonstrations do not suggest that AI scientists can replace human scientists, but they do show that the relevant unit of evaluation should shift from an isolated answer to a trajectory of decisions, artifacts, and evidence. Yet even when an AI scientist produces a plausible new discovery, we often have little understanding of when its reasoning, tool use, or conclusions should be trusted, or which components of the system contributed most to the discovery. A discovery-oriented evaluation framework should therefore decompose the discovery loop into its constituent stages—hypothesis generation, computational or experimental execution, and result interpretation—each producing different artifacts and failing in different ways (Fig. 1). Because no single scalar score can capture these dimensions all together, progress will require a portfolio of evaluations that combine automated scoring, verifiable workflows, and human expert review.

In this Perspective, we argue that evaluating AI scientists must move beyond ever-harder exams toward discovery benchmarks that test whether AI scientists can help produce new and useful scientific results that expand human knowledge. We first examine what exam-style benchmarks measure well and why escalating their difficulty cannot, by itself, certify research capability for discovery (Fig. 1b). We then propose stage-specific evaluation designs for the three stages of the discovery loop, where hypotheses should be verifiable, execution should be increasingly

automated, and interpretation should be robust to uncertainty (Fig. 1c), outlining a roadmap toward closed-loop, end-to-end benchmarks enabled by executable scientific workflows and laboratory automation [30]. Throughout, a central requirement is to distinguish genuine discovery from rediscovery of what is already known. Finally, we identify the data infrastructure this transition requires: evaluations should record negative results, failed attempts, tool calls, intermediate states, and provenance, because these are precisely the data needed to train, audit, and improve reliable AI scientists [31, 32]. Closing this gap will require the scientific community itself to contribute real workflows, executable tasks, and well-documented discovery trajectories. With this infrastructure in place, AI can raise scientific productivity while guarding against shallow reasoning and hidden errors [33], turning today’s benchmark race into a foundation for trustworthy discovery.

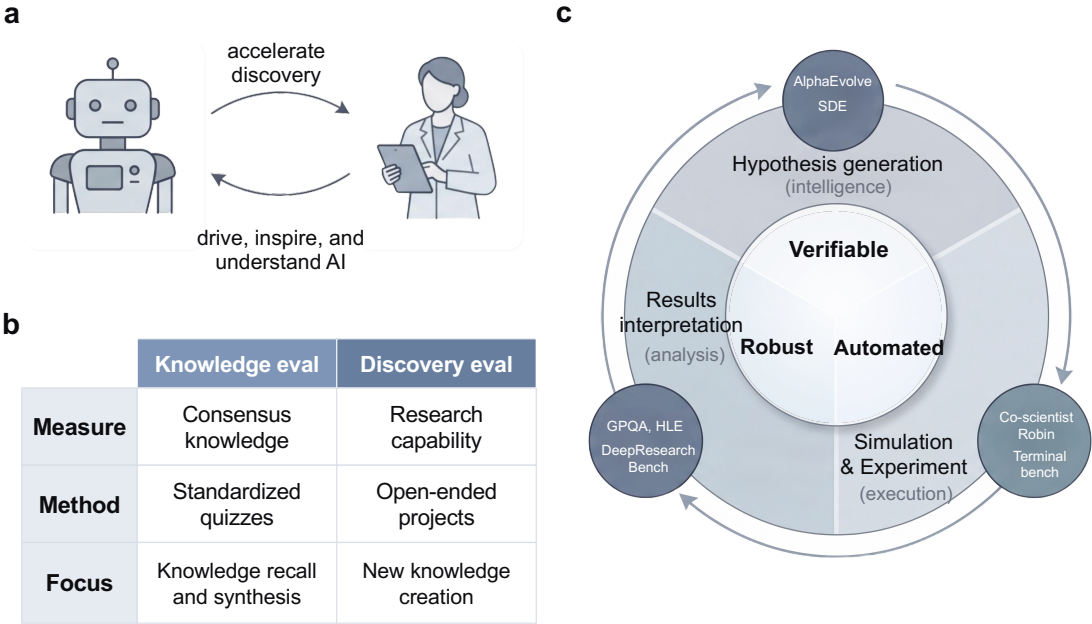


Fig. 1 | From knowledge evaluation to discovery evaluation. **a.** A schematic of the envisioned bidirectional relationship between scientists and LLM-based agents. **b.** Comparison between conventional knowledge evaluation and discovery evaluation. **c.** Conceptual framework for discovery-centric evaluation. Scientific discovery is decomposed into three stages: hypothesis generation, simulation and experiment, and result interpretation, corresponding to intelligence, execution and analysis, respectively. Example work and benchmark families are positioned around this cycle, illustrating how current evaluations probe different parts of the discovery process. At the center are three desirable properties for future benchmarks: they should be verifiable for hypothesis generation, increasingly automated for simulation and experiment execution, and robust for result interpretation.

Main

Moving from knowing science to doing science

Scientists and LLM-based agents are entering a bidirectional relationship, in which agents promise to accelerate discovery and scientists in turn drive, inspire, and help us understand AI and develop better LLMs and agents (Fig.

1a). Evaluation sits at the center of this exchange because benchmarks and models form a mutually reinforcing cycle, where stronger models expose the limits of existing benchmarks and more demanding benchmarks define the future directions for model development. For LLMs in science, this cycle has pushed the benchmark landscape along three axes: breadth, perceptual complexity, and adversarial difficulty, all of which probe what models know rather than what they can do. MMLU first established broad cross-domain academic coverage as a standard reference point [22, 34]; MMMU extended this logic to more complex diagrams, chemical structures, plots, and other expert visual artifacts [23]; GPQA raised the difficulty to PhD-level biology, chemistry, and physics questions that non-experts cannot easily resolve even with web access [24]; and OlympiadBench, HLE, and FrontierScience now push models toward longer chains of mathematical and scientific reasoning and expert-authored answers [25, 26, 35], a direction now embodied by reasoning models trained with reinforcement learning [36, 37]. This design lineage is useful because it removes the heterogeneity of real scientific research. Each benchmark task has a stable input, a sanctioned answer, and a scoring protocol that permits repeated comparison across model release cycles (Fig. 1b). These benchmarks, however, can only measure discriminative problem-solving over established scientific knowledge under constrained input and output formats.

When scores saturate, escalating benchmark difficulty is a straightforward short-term response, but it becomes less sustainable as the frontier moves closer to research practice [38–40]. First, the cost rises sharply: HLE required a global expert pipeline, multi-stage review, and private held-out data, while FrontierScience relies on PhD-level task authors and rubric design [25, 26]. Second, as benchmark questions approach the research frontier, the boundary between “hard but well-posed” and “contestable or under-specified” becomes fragile, especially when grading is executed without verifiable outcomes. Third, hard questions can be mistaken for research competence. A model may excel at deriving an answer in exams while still having no demonstrated ability to design an experiment, choose the right measurement, manage experimental constraints, and, crucially, recognize when the question itself is ill-formed or spot a potential breakthrough in an anomalous observation. Fourth, and most importantly for scientific use, high scores do not certify reliability. HLE reports poor calibration on frontier questions, and broader studies show that larger, more instruction-tuned models may give plausible wrong answers rather than abstain when they should [25, 41, 42]. For scientific discovery, a system’s operating range matters as much as its peak score: knowing when evidence is insufficient, when assumptions fail, and when a result should not yet be trusted is part of the capability in question.

A practical definition of doing science should treat an AI scientist as a policy over a changing epistemic state, rather than as a function from prompt to answer [43, 44]. At any step, the system has background knowledge, prior observations, tools or skills, constraints, and unresolved uncertainty. It proposes a hypothesis or objective, chooses an experiment or simulation, obtains an observation, analyzes the results, updates its state, and decides whether to refine, reject, or extend the hypothesis (Fig. 1c). This formulation connects older robot-scientist work

and the Nobel Turing Challenge to current LLM and agent benchmarks: the essential unit is a loop of proposal, intervention, evidence, and revision, not a single response [45–47]. This framing also separates two capabilities that are often conflated. LLM capability concerns the base model’s ability to retrieve, synthesize, reason, and communicate scientific knowledge. Agent capability concerns whether a system can maintain state, use tools, execute protocols, recover from errors, respect resource and safety constraints, and preserve provenance across many steps. Along this direction, ScienceWorld made this distinction concrete, demonstrating that models which excel at exam-style science questions can fail to complete experiments an elementary-school student could perform, once those experiments must be carried out step by step in an interactive environment [48]. DiscoveryWorld extended this gap to complete discovery cycles of hypothesis formation, experimental action, and analysis [49]. Newer benchmarks such as BLADE, CORE-Bench, LAB-Bench, ScienceAgentBench, ReplicationBench, and SciAgentArena further show that data analysis, reproducibility, protocol planning, research-code execution, and research replication expose failure modes not visible in simple knowledge or reasoning based evaluations [7, 50–54], and scenario-grounded benchmarks have begun to evaluate models at the level of entire research projects that span hypothesis generation, experimental design, and interpretation [55]. A discovery-driven evaluation should therefore score both the final scientific artifact and the trajectory that produced it (Figure 2), for example, whether the system recovers from failed actions, whether the analysis meets its assumptions, whether negative results are correctly incorporated, and whether the claimed discovery is genuinely novel rather than a rediscovery of known results, and empirically supported by the generated evidence.

Measuring the stages of scientific discovery

Scoring both artifacts and trajectories across a full “discovery episode” is demanding, and a practical path is to evaluate each stage of the loop separately before composing them into end-to-end tasks (Fig. 3a). Each stage produces different artifacts, requires different forms of validation, and has different failure modes. Hypothesis generation stresses novelty and verifiability; execution requires reliability and automation; result interpretation demands robustness and evidence grounding.

Hypotheses that can be tested. Hypothesis generation is the most open-ended stage. In an exam, the correct answer is known in advance, while in hypothesis generation, the value of an answer may depend on evidence that does not yet exist. This makes evaluation inherently difficult, because the most interesting hypotheses are often out of distribution, not easily compared to a ground-truth label, and hard to measure with a single score [14]. These difficulties make novelty a natural starting point for hypothesis evaluation [56]. Scientific discovery is valued because it expands what is known; a system that only reconstructs familiar directions may demonstrate scientific knowledge, but not discovery capability. For LLM-based systems, however, apparent novelty may be memorization, as an idea can look original to an expert evaluator while being well represented in the model’s training corpus. Novelty should

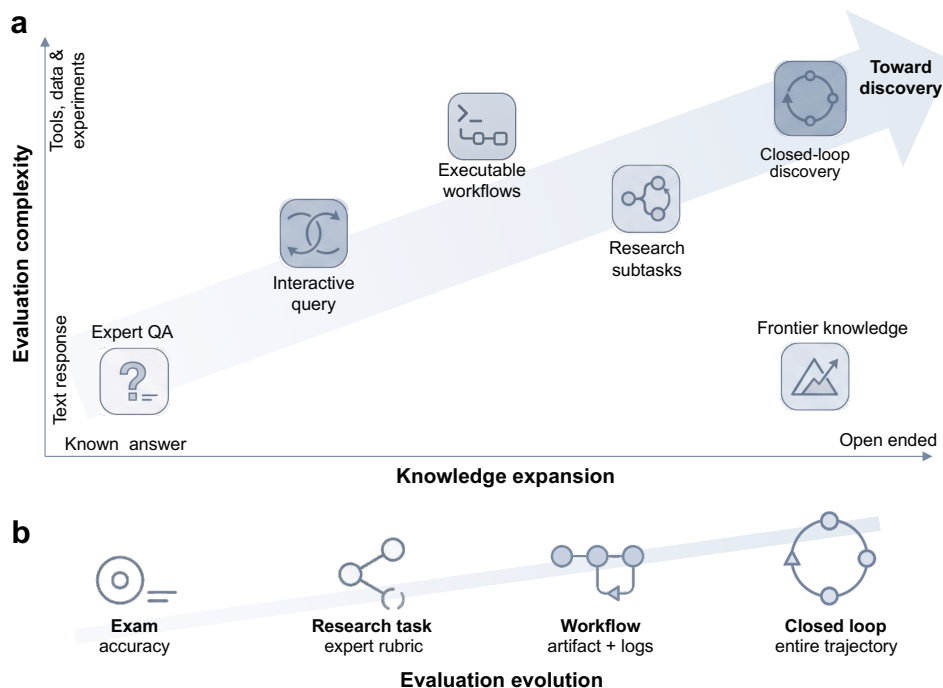


Fig. 2 | Landscape of AI scientist evaluations. **a.** Schematic map of evaluations organized by two dimensions: knowledge expansion, from known-answer tasks to open-ended inquiry, and evaluation complexity, from text responses to evaluations grounded in tools, data and experiments. The shaded trajectory indicates the desired movement toward evaluations that more directly measure discovery-oriented scientific work. **b.** Evaluation methods must evolve in parallel, from exam accuracy on fixed questions to expert-rubric assessment of research tasks, reproducible workflow artifacts and logs, and finally closed-loop evaluation of entire research trajectories. Positions are schematic and indicate the structure of evaluation rather than measured benchmark performance.

therefore be assessed against both the published literature and, where possible, the training data, so that genuine discovery can be distinguished from rediscovery. Yet even genuine novelty is not sufficient on its own. In blinded expert reviews, LLM-generated research ideas were judged more novel than those of human experts, but weaker on feasibility [57]. Useful hypotheses should be diverse as a set, and each should be scientifically plausible, physically realizable, testable, and tied to a measurable objective. Benchmarks should thus treat novelty as necessary but not sufficient, and new ideas must be grounded enough to support computational or experimental verification [58].

Scientific plausibility and physical realizability therefore become essential evaluation criteria. Once a hypothesis passes a basic novelty screen, the relevant question is no longer whether it is interesting in the abstract, but whether the proposed entity can exist, be realized, and be tested under realistic constraints. For example, if an agent proposes a new metal-organic framework for gas separation [59], evaluation should go beyond expert impressions of creativity and check chemical validity, synthetic accessibility, thermodynamic stability [60], and predicted adsorption behavior. If an agent proposes an extension of the Standard Model in high energy physics, the benchmark should verify theoretical self-consistency, such as anomaly cancellation and unitarity, check compatibility with precision measurements and existing collider and cosmological constraints, and ask whether the theory yields falsifiable predictions within reach

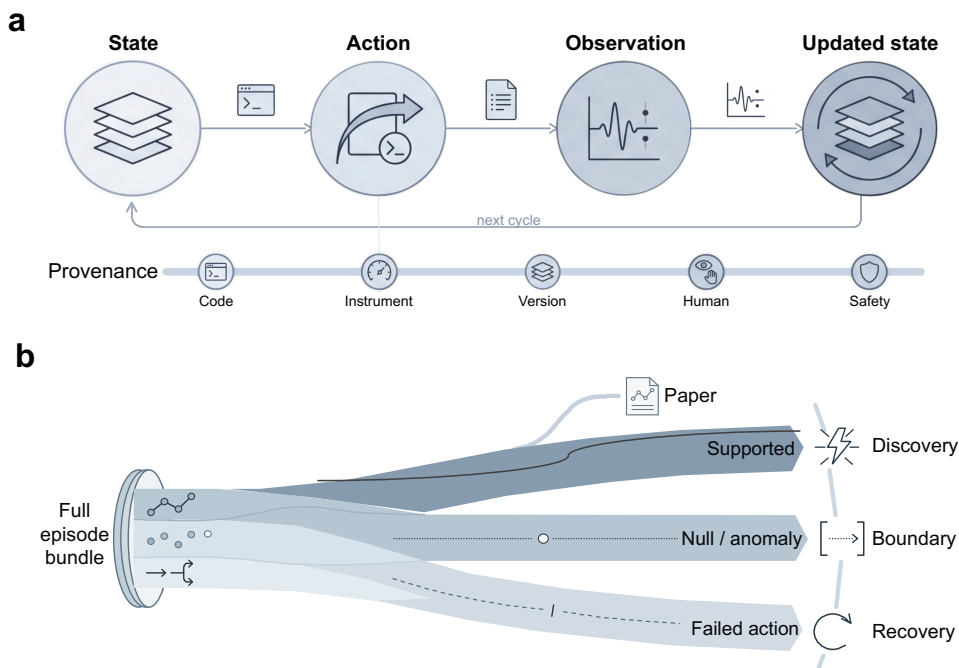


Fig. 3 | Anatomy of a discovery episode. a. A discovery episode can be represented as a structured trajectory from state to action, observation and updated state. The state encodes the current scientific context; the action records the intervention, such as a tool call, computation, experiment, or analysis. The observation captures the resulting evidence. The updated state records how the agent’s beliefs, objectives, or constraints change before the next cycle. A persistent provenance layer links each transition to the code, instruments, versions, human interventions and safety constraints needed for reproducibility and audit. **b.** Full episode bundles preserve signals that are often filtered out of conventional papers. Supported trajectories may become reported discoveries. Null or anomalous outcomes define boundary conditions and failed actions provide recovery information, which are usually left out in publications. Recording all branches turns discovery episodes into reusable evidence for benchmarking, training and evaluating AI scientists.

of current or planned experiments [61]. In both cases, the goal is to move from evaluating the originality of an idea to evaluating whether it can support a concrete computational or experimental test [62].

Hypothesis generation should also be evaluated as an iterative procedure. Human scientists rarely pursue a single hypothesis in isolation; they explore multiple directions in parallel, balance exploration and exploitation, prioritize candidates, and update their beliefs after failure. AI scientists should similarly be assessed on the diversity of their hypotheses, the calibration of their expected outcomes, and the value of information provided by the proposed tests [63, 64]. A collection of moderately novel, feasible, and informative hypotheses may be more scientifically useful than a single bold but untestable claim.

Several benchmark formats are feasible for evaluating emerging hypothesis-generation systems [6, 29]. Retrospective tasks can hide historical discoveries and ask whether agents can recover plausible research trajectories without direct access to the final result—a design that is valid only when the hidden discovery lies outside the model’s training data [65, 66]. Prospective tasks can ask agents to propose candidates in a bounded design space, later

evaluated by simulation or experiment [67]. Online tasks can allow agents to interact with evolving databases or automated workflows. Across all formats, the key requirement is that hypotheses must be connected to verifiable consequences. A discovery benchmark should reward ideas that can survive contact with the world [68].

Execution under real-world constraints. Execution is the stage where a hypothesis becomes an action, and it is the ultimate test of whether that hypothesis holds. Evaluating execution must therefore address two questions. First, can the experiment be carried out with best-practice methods within the given constraints? Second, do its results speak directly to supporting or refuting the hypothesis? Execution spans both the computational and physical realms, as some hypotheses can be tested purely by analyzing, simulating, and modeling data, while others require physical assays. The stage thus encompasses experimental and analytical design, protocol and code writing, instrument and compute infrastructure, and recovery from failures or bugs. Crucially, any evaluation must also reflect the realities of infrastructure, including which instruments are available, how much compute can be used, and what budget the work is allowed to spend. This complexity has left execution an under-evaluated stage of AI for science, yet for the same reasons it carries outsize importance for assessing systems that are increasingly deployed in real-world research.

Purely computational experiments may at first appear to benefit from the same dynamics as software engineering, where deterministic checks such as unit tests have accelerated the evaluation, training, and resulting performance of AI systems [69–71]. They do gain from the possibility of rapid iteration, but they remain experiments, and so stay open-ended, are often agnostic to the process taken, and can yield highly interpretive outcomes. Efforts to constrain such tasks have defined standardized environments and task formats, such as Harbor [72] and Inspect [73], which substantially improve reproducibility and robustness. These formats also make it easier to implement denser rewards across rollouts, so that the approach can be scored alongside the outcome.

Experiments that engage the physical world, such as wet-lab biology, present a fundamentally different challenge. The combined effect of method-agnosticity and resource constraints makes it nearly impossible to design a cleanly evaluable experiment without imposing significant up-front limitations. This difficulty is itself a valuable target for evaluation, as it asks whether an AI scientist can design not only the ideal experiment in the abstract, but the optimal one to run with the resources at hand. Some groups have evaluated experimental design through focused projects, such as improving protein yield in a cell-free expression system [74], cloning and protein modification [13], or drug repurposing [12], but building a dedicated benchmark around wet-lab execution remains an open challenge.

A further source of variability is human execution itself. No scientist is infallible in the lab, and no experiment performs identically in all hands or all facilities. This variability may appear to be a roadblock to evaluating physical execution, but it also defines an opportunity, namely to ask whether an AI scientist can design experiments and protocols that are robust to the executor. Automated laboratories, first popularized in the life sciences and now expanding into chemistry and materials, are the often-cited solution, as they are more consistent, faster, and more

scalable [10, 20]. Yet they do not capture the many experiments that resist automation because they are not iterative or scalable, or because their parameters vary too widely, and automated platforms are themselves not yet flexible enough. Here, humanoid robots (sometimes referred to as “physical AI”) are a promising direction, as they can in principle reproduce the full range of human laboratory actions and even operate under conditions that are unsafe for human scientists.

These constraints raise the central question of how execution should be scored, and in particular whether evaluation should reward the process, the outcome, or both. The key confounder is method-agnosticity. Human experts frequently disagree on the best analysis or experiment for a given hypothesis, whether in the choice of method and protocol or merely in reagents and parameters. There is often a wide breadth of valid approaches and yet a fine line between valid and invalid, and even an invalid approach may reach the same qualitative conclusion. Evaluating execution therefore requires positioning the measurement on a spectrum from opinionated to accommodating. An opinionated evaluation accepts a single approach as correct, even when that does not reflect technical reality, whereas an accommodating one admits multiple valid approaches, though being truly comprehensive is often untenable. An opinionated evaluation can be easier to build, because it relies on discrete, easily scored outcomes and leaves the correctness of the path taken implicit in the rewarded outcome. How the outcome is defined sets the degree of opinionation. Requiring an exact numerical value is highly opinionated, whereas requiring only that a target gene of interest appear among the top ten differentially expressed genes is more permissive, since it tolerates variation in the underlying approach.

The same spectrum can be modulated at the task level. A task may supply a high-level plan that constrains the approach to the outcome [75], or specify the procedure in greater detail [76]. Such scaffolding reduces the method-agnosticity problem, but it makes the task easier, forecloses equally valid alternatives, and risks measuring instruction-following rather than scientific ability. It also departs from the real-world use of AI scientists, which should be expected to choose the right approach autonomously. The alternative is to pose the task in an open-ended form, where the correctness of the path is again implicit in the rewarded outcome [77–80]. Here too the expected outcome can be defined narrowly, or admit several equally valid results. Evaluating the approach in addition to the outcome demands more complex assessment, but it enables denser reward signals and partial credit, for example through comprehensive rubrics and LLM judges, which can themselves be tuned along the opinionated-to-accommodating spectrum.

Interpretation with uncertainty. Result interpretation is the stage where computational or experimental outputs become scientific evidence. At first glance, this stage may appear easier to evaluate than hypothesis generation because the system is given data rather than asked to invent a new direction. In practice, however, interpretation is often where discovery benchmarks become most fragile. Scientific measurements are rarely self-explanatory, as

they are shaped by instrument calibration, preprocessing choices, statistical assumptions, experimental controls, and domain-specific conventions. A benchmark that asks only for a final conclusion risks rewarding plausible narratives rather than calibrated scientific judgment. A useful interpretation benchmark should therefore pair each experiment or simulation with best-practice analysis methods appropriate to the data [7, 50]. An AI scientist should not merely summarize results, but choose or apply analyses whose assumptions are justified, for example by checking sample size, control groups, noise, convergence criteria, uncertainty estimates, and possible confounders. In computational settings, this may include verifying that a simulation has converged or that the chosen model is valid for the regime being studied. In experimental settings, it may require recognizing batch effects, instrument drift, contamination, missing metadata, or measurements near the detection limit.

Interpretability and calibration to scientific instruments are especially important. Instruments such as NMR spectrometers, mass spectrometers, microscopes, diffractometers, and electrochemical workstations do not produce direct scientific conclusions. Instead, they produce signals that require domain-specific interpretation. For example, an AI scientist analyzing NMR data should be able to distinguish expected peaks from solvent signals, impurities, side products, baseline artifacts, or referencing errors—a capability that frontier models have recently begun to demonstrate at the level of dedicated spectroscopy software [81]. If the NMR evidence conflicts with mass spectrometry or another characterization method, the system should not simply choose the more convenient result. Rather, it should explain the conflict, assess which interpretations remain plausible, and propose a follow-up measurement that would reduce uncertainty.

Since interpretation depends on how evidence is produced, analyzed, and combined, this stage should be evaluated as an evidentiary process rather than a final-answer task. A strong benchmark should evaluate the full evidentiary chain, from extraction of relevant signals in raw data to calibration of uncertainty, comparison across multiple measurements, and synthesis into a justified conclusion. A capable AI scientist should be able to orchestrate evidence across modalities, separate observation from interpretation, and state clearly what the data do and do not support. Recent agentic frameworks that subject free-form hypotheses to sequential statistical falsification, with explicit control of type-I error, illustrate how such discipline can be automated [82]. In many cases, the best answer will not be a decisive claim, but a robust judgment that neither overstates weak evidence nor discounts strong evidence, acknowledging that several explanations remain consistent with the data.

This process-oriented view also exposes a central challenge, namely that interpretation often requires human-defined evaluation. Unlike closed-ended exams, many scientific results do not have a single canonical answer, especially when the data are noisy, incomplete, or contradictory. This ambiguity grows as an AI scientist approaches cutting-edge research, where the observations most likely to signal a knowledge breakthrough are precisely those for which no established interpretation yet exists [83]. Expert rubrics are therefore necessary to assess whether

an interpretation is statistically valid, instrument-aware, and appropriately cautious. Rubric-based “deep research” evaluations already take this approach for literature-grounded synthesis tasks [26], and extending it from curated sources to raw experimental data is the natural next step. These rubrics should reward not only correctness, but also uncertainty calibration, identification of failure modes, reproducibility of the analysis pipeline, and quality of the recommended follow-up action. In this sense, interpretation benchmarks should measure whether an AI scientist can reason with evidence under uncertainty, rather than merely generate a convincing post-hoc explanation. When performed well, interpretation closes the discovery loop, where rigorous analysis of the evidence determines whether a hypothesis should be refined, rejected, or pursued further.

Roadmap to closed-loop, end-to-end discovery benchmarks

The stage-wise evaluations of hypothesis generation, execution, and interpretation each measure distinct competencies, but scientific discovery ultimately emerges only from their composition and the feedback loop that connects them. A hypothesis is turned into an experiment and executed, the resulting evidence is interpreted, and the interpretation feeds back to inform the next hypothesis, closing the loop where it began. Each of the three stage-wise criteria—verifiable hypotheses, automated execution, and robust interpretation—remains necessary when evaluating the discovery process as a whole, but together they are still not sufficient. Any closed-loop, end-to-end evaluation framework for AI scientists must additionally satisfy several properties that arise only once the loop turns.

First, within a single episode, scoring must assign credit across its stages, since the end-to-end result reflects their interaction rather than any stage in isolation. A strong, verifiable hypothesis can still be wasted by poor execution, and a flawed interpretation can turn a well-run experiment into a false conclusion. A failure can therefore not be isolated to one stage without examining the full trajectory. A closed-loop benchmark must accordingly score the trajectory itself—attributing the outcome to the stage responsible for it—rather than judging the final result alone.

Second, evaluation must extend across episodes, not only within them. Two agents that behave identically on a single problem can differ entirely in what they carry forward, namely whether they exploit what earlier episodes established—transferring insight and avoiding dead ends already mapped—and whether they explore what those episodes left open, probing the untried regions most likely to reward attention. What a closed-loop benchmark must evaluate is not a fixed strategy but how well an agent steers its own discovery process across episodes—accumulating scientific experience and deploying it in both directions—so that the search grows better informed and more efficient with each new episode.

Third, a discovery made through this iterative process must still survive replication, the standard by which science separates real findings from artifacts. A closed-loop benchmark can test this rather than trust the agent’s report [53, 84]. Because the experiments run in a live computational environment or physical laboratory, the finding can

be replicated independently and counted only if it holds. In a computational setting this is already practical, as the verification plan can be executed independently, cheaply, and automatically, and in the physical world automated laboratories are steadily closing the gap [10, 20, 85].

At scale, the loop breaks at the execution stage. An agent can still generate a hypothesis and interpret the data once it exists, but what it cannot do is run the experiment that would test it when that means a wide-field astronomical survey, a multi-year clinical trial, a decade-long climate field campaign, or a new particle accelerator. Hypothesis and interpretation scale, but execution does not, and so the loop cannot close. A minimum closed-loop benchmark must therefore stay in the regime where execution is containable—a bounded, well-characterized problem with an automatically computable objective, run in a controlled environment where a hypothesis can be tested at reasonable time and cost.

Closing the data gap: from static corpora to “discovery episodes”

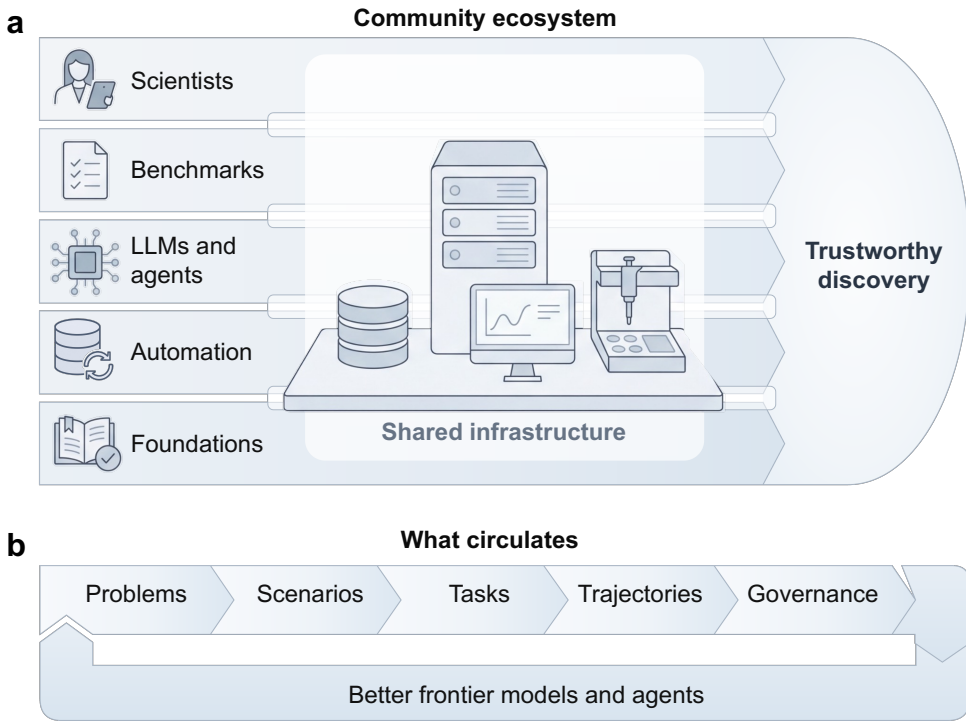


Fig. 4 | A community ecosystem for measuring AI scientists. **a.** Discovery-oriented evaluation requires a shared infrastructure that connects scientists, benchmarks, LLMs and agents, automation systems and foundational institutional support. Scientists contribute real problems and expertise; benchmarks translate them into measurable tasks; LLMs and agents attempt the tasks and expose capabilities and failure modes; automation infrastructure captures data, experiments and trajectories; and foundations provide standards, credit and oversight. Together, these layers support trustworthy discovery rather than isolated benchmark performance. **b.** The ecosystem is sustained by what circulates through it: problems, scenarios, tasks, trajectories and governance mechanisms. As these elements are refined and reused, they feed back into better frontier models and agents, creating a community loop for improving both evaluation and scientific discovery.

The limiting data problem for AI scientists is not only the amount of scientific text available, but the form in which scientific work is recorded. Closed-ended benchmarks usually compress evaluation into (question $\rightarrow$ answer), while papers compress research into a retrospective narrative of (problem $\rightarrow$ evidence $\rightarrow$ claim). Both formats are valuable for human communication, but they discard much of the information needed by agents that must act over time. This limitation is especially acute in reinforcement learning, in which improving a scientific agent through interaction requires reward signals that are denser than a final publication or a pass-fail benchmark score [86]. Useful signals include intermediate feedback on whether a tool call was valid, an experiment was feasible, a measurement was interpretable, or a recovery action moved the system closer to a reliable result. An AI scientist therefore needs trajectories of scientific research: the initial state of knowledge, candidate hypotheses considered and rejected, tool calls, code edits, simulation inputs, instrument settings, intermediate observations, failed attempts, recovery actions, and the evolving rationale that connects one decision to the next. Recent agent benchmarks in scientific environments have begun to expose these requirements, but most still evaluate bounded tasks rather than complete discovery histories [7, 48, 49]. The data object needed for training and evaluation is therefore a “discovery episode”: an AI-ready record of (state $\rightarrow$ action $\rightarrow$ observation $\rightarrow$ updated state) repeated until a scientific claim is supported, abandoned, or revised (Figure 3a). Existing standards and infrastructures already point in this direction. FAIR principles emphasize findable, accessible, interoperable, and reusable scientific data [87]. Efforts such as the Open Reaction Database and the Materials Project, workflow platforms like AiiDA and FireWorks, and self-driving laboratories have already illustrated the value of standardized reaction conditions, parameter histories, instrument metadata, and outcome records that preserve provenance at scale [30, 88–92].

Another data gap is that the scientific record is biased toward what worked. Published papers over-represent positive, clean, and coherent results that support their narratives, while failed synthesis, abandoned hypotheses, debugging steps, and inconclusive measurements are rarely preserved in reusable form [31, 93]. This omission serves human efficiency, but it discards exactly what agents could use. Unlike humans, agents can ingest the full record of failures at scale, learning from data that no individual scientist could ever read through. Negative results are not merely missing publications, but training signals: they define the boundary of a hypothesis, reveal hidden constraints, teach when to stop, and reduce the tendency to rediscover known dead ends. A robust AI scientist should learn not only which action produced a reported success, but also which nearby actions failed, which assumptions were violated, which measurements were noisy, and which recovery strategies restored progress. The next step is to produce large-scale “discovery episodes” that include experiments and simulations, intermediate analysis, negative and ambiguous outcomes, decision rationales, software environments, versioned protocols, and human interventions (Figure 3b). Capturing these episodes need not add to scientists’ burden: self-driving laboratories and workflow engines can record them automatically as a by-product of normal operation [30, 90], and journals could extend FAIR-style requirements from final datasets to discovery trajectories, crediting the researchers who contribute them

[87]. Such records would let AI scientists train on the full distribution of scientific practice rather than the polished subset that survives publication. They would also make evaluation more auditable, where a claimed discovery could be inspected not just as an endpoint, but as a reproducible path through evidence, uncertainty, failure, and revision.

Discussion

Measuring AI scientists is a community undertaking that no single group can accomplish alone. It requires a shared infrastructure that connects scientific understanding, benchmark construction, computational validation, and experimental verification (Figure 4a). Scientists must contribute real workflows, meaningful objectives, and representative failure cases, benchmark builders must convert these workflows into scalable and reproducible evaluation environments, and AI researchers and frontier labs must optimize agents against these environments and expose their failure modes. If this loop succeeds, improved agents will flow back into science, helping researchers generate hypotheses, execute workflows, interpret results, and create new discovery episodes for the next generation of systems.

Discovery benchmarks also cannot remain static. Scientific agents such as Kosmos, Claude Science, and MIRA are already moving from research prototypes to deployed platforms that couple literature search, data analysis, and long-horizon task execution behind a single interface, and their rapid release cycle ensures that any fixed benchmark will measure a moving target[94–96]. The demands placed on AI scientists will evolve as models, agents, tools, and lab automation improve, and what counts as a meaningful scientific contribution will shift as larger design spaces, faster experimental cycles, and new forms of human-AI collaboration are unlocked. The criteria for “good science” must evolve accordingly, as success in discovery depends on producing ideas, workflows, and interpretations that are novel, testable, reproducible, useful, and safe. As AI scientists explore more possibilities than human researchers can manually inspect, additional criteria will grow in importance, including calibration under uncertainty, cost-adjusted value of information, robustness to failed experiments, interpretability of decisions, and the ability to ask for human oversight. The goal should therefore not be one final benchmark for scientific discovery, but a living measurement ecosystem in which tasks, metrics, and reporting standards evolve with scientific practice while comparability across systems is preserved, moving the benchmark target from isolated correctness toward disciplined scientific judgment (Figure 4b).

Reliability must sit at the center of this evolving ecosystem. AI systems have begun to produce striking scientific and mathematical results, including formally verified progress on long-standing Erdős problems [97, 98] and olympiad-level mathematical reasoning [99]. At the same time, some celebrated examples of “solving” open problems in fact surfaced solutions already published in the literature, a reminder that distinguishing discovery from rediscovery is itself part of the measurement problem [100]. A single breakthrough at the boundary of human knowledge, however impressive, is different from a reliable discovery engine. The central evaluation question is therefore not whether

an AI scientist can occasionally solve a hard problem, but whether it can repeatedly generate useful progress under controlled, auditable, and reproducible conditions.

Closely related to reliability is safety. Scientific discovery can create powerful capabilities, including in chemistry, materials, energy, and life sciences. Benchmarks that reward only speed, novelty, or objective improvement could unintentionally incentivize unsafe behavior, such as proposing hazardous experiments, optimizing poorly specified objectives, ignoring uncertainty, bypassing human oversight, or generating dual-use knowledge without appropriate safeguards [101, 102]. Safety should therefore be built into benchmark design rather than appended afterward. A discovery benchmark should specify allowed design spaces, forbidden objectives, safety constraints, human-review points, provenance requirements, and conditions under which the agent should stop or ask for help. In some cases, the safest and most scientifically responsible action is not to proceed, but to recognize that the evidence, protocol, or objective is insufficiently constrained.

Beyond evaluation, science may become one of the most important environments for developing more capable AI. Scientific discovery naturally provides what many current training settings lack, namely long-horizon objectives, interaction with external tools, uncertain observations, delayed feedback, and consequences that can be checked against reality [103]. Closed-loop discovery environments could therefore serve not only as benchmarks but also as learning environments, in which the discovery episodes described above become training signal rather than archive. In this sense, measuring AI scientists may help unlock more general forms of intelligence, because pushing science forward requires reasoning, planning, acting, observing, and revising beliefs in contact with the world [47].

Together, these considerations define the next phase of AI scientists evaluation. Exam-style benchmarks remain useful for measuring scientific knowledge, but they are only one component of a broader measurement ecosystem. Stage-specific benchmarks test whether knowledge can be converted into hypotheses, executable workflows, and calibrated interpretations. Closed-loop benchmarks test whether these capabilities can be composed into reliable discovery over time. The long-term goal is not acceleration alone, nor the replacement of scientific judgment with automated output, but AI scientists that can participate responsibly in the scientific discovery process. The field should therefore move from asking whether AI can answer harder scientific questions to asking whether AI can help create trustworthy scientific breakthroughs. That transition will require community effort, evolving standards, and benchmarks that look less like exams and more like science itself.

Acknowledgment

C.D. thanks Frances Arnold for the conceptualization and discussions of this work via private communications. P.J. supported in part by Defense Advanced Research Projects Agency’s (DARPA) SciFy program (Agreement No. HR00112520300),

Author Contributions

Y.D., S.D., J.L., and C.D. wrote the initial draft. All authors participate in the conceptualization of this project and reviewed and edited the final draft.

Competing interests

The authors declare that they have no competing financial interests at this time. P.J. has an outside interest in the Allen Institute for Artificial Intelligence. This interest has been disclosed to the University of Arizona and reviewed in accordance with its conflict of interest policies.

References

1. Wang, H. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60. DOI: 10.1038/s41586-023-06221-2 (2023).
2. Skarlinski, M. D. *et al.* Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*. arXiv: 2409.13740 [cs.CL] (2024).
3. Aygün, E. *et al.* An AI system to help scientists write expert-level empirical software. *Nature* **654**, 909–916. DOI: 10.1038/s41586-026-10658-6 (2026).
4. Romera-Paredes, B. *et al.* Mathematical discoveries from program search with large language models. *Nature* **625**, 468–475. DOI: 10.1038/s41586-023-06924-6 (2024).
5. Novikov, A. *et al.* Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*. DOI: 10.48550/arXiv.2506.13131 (2025).
6. Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., *et al.* Accelerating scientific discovery with Co-Scientist. *Nature*. DOI: 10.1038/s41586-026-10644-y (2026).
7. Chen, Z. *et al.* *ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery* in *International Conference on Learning Representations* **2025** (2025), 96934–96990. DOI: 10.48550/arXiv.2410.05080.
8. Majumder, B. P. *et al.* *Discoverybench: Towards data-driven discovery with large language models* in *International Conference on Learning Representations* **2025** (2024), 4556–4579. DOI: 10.48550/arXiv.2407.01725.
9. M. Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **6**, 525–535. DOI: 10.1038/s42256-024-00832-8 (2024).

10. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578. DOI: 10.1038/s41586-023-06792-0 (2023).
11. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* **646**, 716–723. DOI: 10.1038/s41586-025-09442-9 (2025).
12. Ghareeb, A. E. *et al.* A multi-agent system for automating scientific discovery. *Nature*. DOI: 10.1038/s41586-026-10652-y (2026).
13. Huang, K. *et al.* Autonomous biomedical research with an artificial intelligence agent. *Science*, eadz4351. DOI: 10.1126/science.adz4351 (2026).
14. Du, Y. *et al.* Accelerating scientific discovery with autonomous goal-evolving agents. *arXiv preprint arXiv:2512.21782*. DOI: 10.48550/arXiv.2512.21782 (2025).
15. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589. DOI: 10.1038/s41586-021-03819-2 (2021).
16. Abramson, J., Adler, J., Dunger, J., *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500. DOI: 10.1038/s41586-024-07487-w (2024).
17. Davies, A. *et al.* Advancing mathematics by guiding human intuition with AI. *Nature* **600**, 70–74. DOI: 10.1038/s41586-021-04086-x (2021).
18. Gómez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science* **4**, 268–276. DOI: 10.1021/acscentsci.7b00572 (2018).
19. Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology* **37**, 1038–1040. DOI: 10.1038/s41587-019-0224-x (2019).
20. Szymanski, N. J. *et al.* An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature* **624**, 86–91. DOI: 10.1038/s41586-023-06734-w (2023).
21. Jansen, P., Clark, P., Downey, D. & Weld, D. S. *Generating Literature-Driven Scientific Theories at Scale in Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Liakata, M., Moreira, V. P., Zhang, J. & Jurgens, D.) (Association for Computational Linguistics, San Diego, California, United States, July 2026), 14677–14698. DOI: 10.18653/v1/2026.acl-long.669.
22. Hendrycks, D., Burns, C., Basart, S., *et al.* *Measuring Massive Multitask Language Understanding in International Conference on Learning Representations* (2021). DOI: 10.48550/arXiv.2009.03300.
23. Yue, X., Ni, Y., Zhang, K., *et al.* *MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024). DOI: 10.1109/CVPR52733.2024.00913.

24. Rein, D. *et al.* GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*. arXiv: 2311.12022 [cs.AI] (2023).
25. Center for AI Safety, Scale AI & HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature* **649**, 1139–1146. DOI: 10.1038/s41586-025-09962-4 (2026).
26. Wang, M. *et al.* FrontierScience: Evaluating AI’s Ability to Perform Expert-Level Scientific Tasks. *arXiv preprint arXiv:2601.21165*. DOI: 10.48550/arXiv.2601.21165. arXiv: 2601.21165 [cs.AI] (2026).
27. Box, G. E. P. Science and statistics. *Journal of the American Statistical Association* **71**, 791–799. DOI: 10.1080/01621459.1976.10480949 (1976).
28. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* 2nd ed. (MIT Press, Cambridge, MA, 2018).
29. Lu, C. *et al.* Towards end-to-end automation of AI research. *Nature* **651**, 914–919. DOI: 10.1038/s41586-026-10265-5 (2026).
30. Volk, A. A. & Abolhasani, M. Performance metrics to unleash the power of self-driving labs in chemistry and materials science. *Nature Communications* **15**, 1378. DOI: 10.1038/s41467-024-45569-5 (2024).
31. Fanelli, D. Negative results are disappearing from most disciplines and countries. *Scientometrics* **90**, 891–904. DOI: 10.1007/s11192-011-0494-7 (2012).
32. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76. DOI: 10.1038/nature17439 (2016).
33. Messeri, L. & Crockett, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58. DOI: 10.1038/s41586-024-07146-0 (2024).
34. Wang, Y., Ma, X., Zhang, G., *et al.* MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *arXiv preprint arXiv:2406.01574*. DOI: 10.48550/arXiv.2406.01574 (2024).
35. He, C., Luo, R., Bai, Y., *et al.* OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. *arXiv preprint arXiv:2402.14008*. DOI: 10.48550/arXiv.2402.14008 (2024).
36. OpenAI. OpenAI o1 System Card. *arXiv preprint arXiv:2412.16720*. DOI: 10.48550/arXiv.2412.16720 (2024).
37. Guo, D., Yang, D., Zhang, H., *et al.* DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638. DOI: 10.1038/s41586-025-09422-z (2025).
38. Reuel, A. *et al.* *BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices in NeurIPS Datasets and Benchmarks Track* (2024). DOI: 10.48550/arXiv.2411.12990.
39. Recht, B., Roelofs, R., Schmidt, L. & Shankar, V. *Do ImageNet Classifiers Generalize to ImageNet?* in *International Conference on Machine Learning* **2019** (2019), 5389–5400. DOI: 10.48550/arXiv.1902.10811.

40. Gadre, S. Y., Ilharco, G., Fang, A., *et al.* *DataComp: In search of the next generation of multimodal datasets in Advances in Neural Information Processing Systems* **36** (2023). DOI: 10.48550/arXiv.2304.14108.
41. Zhou, L., Schellaert, W., Martínez-Plumed, F., *et al.* Larger and more instructable language models become less reliable. *Nature* **634**, 61–68. DOI: 10.1038/s41586-024-07930-y (2024).
42. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630. DOI: 10.1038/s41586-024-07421-0 (2024).
43. Kaelbling, L. P., Littman, M. L. & Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* **101**, 99–134. DOI: 10.1016/S0004-3702(98)00023-X (1998).
44. Narayanan, S., Braza, J. D., *et al.* Aviary: training language agents on challenging scientific tasks. *arXiv preprint arXiv:2412.21154*. DOI: 10.48550/arXiv.2412.21154 (2024).
45. King, R. D., Whelan, K. E., Jones, F. M., *et al.* Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **427**, 247–252. DOI: 10.1038/nature02236 (2004).
46. King, R. D., Rowland, J., Oliver, S. G., *et al.* The Automation of Science. *Science* **324**, 85–89. DOI: 10.1126/science.1165620 (2009).
47. Kitano, H. Nobel Turing Challenge: creating the engine for scientific discovery. *npj Systems Biology and Applications* **7**, 29. DOI: 10.1038/s41540-021-00189-3 (2021).
48. Wang, R., Jansen, P., Côté, M.-A. & Ammanabrolu, P. *ScienceWorld: Is your Agent Smarter than a 5th Grader?* in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (2022), 11279–11298. DOI: 10.18653/v1/2022.emnlp-main.775.
49. Jansen, P., Côté, M.-A., Khot, T., *et al.* DISCOVERYWORLD: A Virtual Environment for Developing and Evaluating Automated Scientific Discovery Agents. *arXiv preprint arXiv:2406.06769*. DOI: 10.48550/arXiv.2406.06769 (2024).
50. Gu, K., Shang, R., Jiang, R., *et al.* BLADE: Benchmarking Language Model Agents for Data-Driven Science. *arXiv preprint arXiv:2408.09667*. DOI: 10.48550/arXiv.2408.09667 (2024).
51. Siegel, Z. S., Kapoor, S., Nadgir, N., Stroebel, B. & Narayanan, A. CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark. *arXiv preprint arXiv:2409.11363*. DOI: 10.48550/arXiv.2409.11363 (2024).
52. Laurent, J. M., Janizek, J. D., Ruzo, M., *et al.* LAB-Bench: Measuring Capabilities of Language Models for Biology Research. *arXiv preprint arXiv:2407.10362*. DOI: 10.48550/arXiv.2407.10362 (2024).
53. Ye, C. *et al.* *ReplicationBench: Can AI Agents Replicate Astrophysics Research Papers?* 2025. arXiv: 2510.24591 [cs.CL].

54. Liu, T. *et al.* Benchmarking AI Agents for Addressing Scientific Challenges Across Scales. *arXiv preprint arXiv:2606.12736*. DOI: 10.48550/arXiv.2606.12736 (2026).
55. Song, Z., Lu, J., Du, Y., *et al.* Evaluating Large Language Models in Scientific Discovery. *arXiv preprint arXiv:2512.15567*. DOI: 10.48550/arXiv.2512.15567 (2025).
56. Agarwal, D. *et al.* Autodiscovery: Open-ended scientific discovery via bayesian surprise. *Advances in Neural Information Processing Systems* **38**, 25181–25219. DOI: 10.48550/arXiv.2507.00310 (2025).
57. Si, C., Yang, D. & Hashimoto, T. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *arXiv preprint arXiv:2409.04109*. arXiv: 2409.04109 [cs.CL] (2024).
58. Du, Y., Jamasb, A. R., Guo, J., *et al.* Machine learning-aided generative molecular design. *Nature Machine Intelligence* **6**, 589–604. DOI: 10.1038/s42256-024-00843-5 (2024).
59. Yao, Z. *et al.* Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence* **3**, 76–86. DOI: 10.1038/s42256-020-00271-1 (2021).
60. Riebesell, J., Goodall, R. E. A., *et al.* A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence* **7**. DOI: 10.1038/s42256-025-01055-1 (2025).
61. Particle Data Group, Navas, S., *et al.* Review of Particle Physics. *Physical Review D* **110**, 030001. DOI: 10.1103/PhysRevD.110.030001 (2024).
62. Duan, C., Du, Y., Jia, H. & Kulik, H. J. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. *Nature Computational Science* **3**, 1045–1055. DOI: 10.1038/s43588-023-00563-7 (2023).
63. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE* **104**, 148–175. DOI: 10.1109/JPROC.2015.2494218 (2016).
64. Yang, J., Lal, R. G., Bowden, J. C., *et al.* Active learning-assisted directed evolution. *Nature Communications* **16**, 714. DOI: 10.1038/s41467-025-55987-8 (2025).
65. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98. DOI: 10.1038/s41586-019-1335-8 (2019).
66. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85. DOI: 10.1126/science.1165893 (2009).
67. Johnston, K. E. *et al.* A combinatorially complete epistatic fitness landscape in an enzyme active site. *Proceedings of the National Academy of Sciences* **121**, e2400439121. DOI: 10.1073/pnas.2400439121 (2024).
68. Popper, K. R. *The Logic of Scientific Discovery* DOI: 10.4324/9780203994627 (Hutchinson, London, 2002).
69. Jimenez, C. E. *et al.* SWE-bench: Can Language Models Resolve Real-World GitHub Issues? arXiv:2310.06770 [cs]. Nov. 2024. DOI: 10.48550/arXiv.2310.06770.

70. Chan, J. S., Chowdhury, N., Jaffe, O., *et al.* MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. *arXiv preprint arXiv:2410.07095*. DOI: 10.48550/arXiv.2410.07095 (2024).
71. Wijk, H., Lin, T., Becker, J., *et al.* RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. *arXiv preprint arXiv:2411.15114*. DOI: 10.48550/arXiv.2411.15114 (2024).
72. Harbor Framework Team. *Harbor: A framework for evaluating and optimizing agents and models in container environments* version v0.16.1. 2026. DOI: 10.5281/zenodo.20953922.
73. AI Security Institute, U. *Inspect AI: Framework for Large Language Model Evaluations* May 2024.
74. Smith, A. A. *et al.* *Using a GPT-5-driven autonomous lab to optimize the cost and titer of cell-free protein synthesis* en. ISSN: 2692-8205 Pages: 2026.02.05.703998 Section: New Results. Feb. 2026. DOI: 10.64898/2026.02.05.703998.
75. EdisonScientific. *BixBench_Hypothesis* https://huggingface.co/datasets/EdisonScientific/bixbench_hypothesis. 2026.
76. Mitchener, L. *et al.* *BixBench: a Comprehensive Benchmark for LLM-based Agents in Computational Biology* arXiv:2503.00096 [q-bio.QM]. Oct. 2025. DOI: 10.48550/arXiv.2503.00096.
77. Merrill, M. A. *et al.* *Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces* 2026. arXiv: 2601.11868 [cs.SE].
78. Li, J. H. & Ho, A. J. *GeneBench-Pro: Evaluating Multistage Statistical Reasoning in Genomics, Quantitative Biology, and Translational Biomedicine* en. ISSN: 2692-8205 Pages: 2026.06.29.735386 Section: New Results. July 2026. DOI: 10.64898/2026.06.29.735386.
79. Terminal-Bench Science Team. *Terminal-Bench Science: Evaluating AI Agents on Computational Workflows in the Natural Sciences* Feb. 2026.
80. Tian, M., Gao, L., Zhang, S. D., *et al.* SciCode: A Research Coding Benchmark Curated by Scientists. *arXiv preprint arXiv:2407.13168*. DOI: 10.48550/arXiv.2407.13168 (2024).
81. Anthropic. *Making Claude a chemist* <https://www.anthropic.com/research/making-claude-a-chemist>.
82. Huang, K. *et al.* Automated Hypothesis Validation with Agentic Sequential Falsifications. *arXiv preprint arXiv:2502.09858*. arXiv: 2502.09858 [cs.AI] (2025).
83. Dillmann, S., Martínez-Galarza, J. R., *et al.* Representation learning for time-domain high-energy astrophysics: Discovery of extragalactic fast X-ray transient XRT 200515. *Monthly Notices of the Royal Astronomical Society* **537**, 931–955. DOI: 10.1093/mnras/stae2808 (2025).
84. Starace, G., Jaffe, O., Sherburn, D., *et al.* PaperBench: Evaluating AI’s Ability to Replicate AI Research. *arXiv preprint arXiv:2504.01848*. DOI: 10.48550/arXiv.2504.01848 (2025).

85. Burger, B., Maffettone, P. M., Gusev, V. V., *et al.* A mobile robotic chemist. *Nature* **583**, 237–241. DOI: 10.1038/s41586-020-2442-2 (2020).
86. Lightman, H., Kosaraju, V., Burda, Y., *et al.* Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*. DOI: 10.48550/arXiv.2305.20050 (2023).
87. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018. DOI: 10.1038/sdata.2016.18 (2016).
88. Kearnes, S. M. *et al.* The Open Reaction Database. *Journal of the American Chemical Society* **143**, 18820–18826. DOI: 10.1021/jacs.1c09820 (2021).
89. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002. DOI: 10.1063/1.4812323 (2013).
90. Huber, S. P., Zoupanos, S., Uhrin, M., *et al.* AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Scientific Data* **7**, 300. DOI: 10.1038/s41597-020-00638-4 (2020).
91. Jain, A. *et al.* FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurrency and Computation: Practice and Experience* **27**, 5037–5059. DOI: 10.1002/cpe.3505 (2015).
92. Barroso-Luque, L., Shuaibi, M., Fu, X., *et al.* Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models. *arXiv preprint arXiv:2410.12771*. DOI: 10.48550/arXiv.2410.12771 (2024).
93. Duan, C., Janet, J. P., Liu, F., Nandy, A. & Kulik, H. J. Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models. *Journal of Chemical Theory and Computation* **15**, 2331–2345. DOI: 10.1021/acs.jctc.9b00057 (2019).
94. Mitchener, L. *et al.* Kosmos: An AI Scientist for Autonomous Discovery. *arXiv: 2511.02824 [cs.AI]* (2025).
95. Anthropic. *Claude Science, an AI workbench for scientists* <https://www.anthropic.com/news/claude-science-ai-workbench>.
96. Deep Principle. *Mira: An AI Scientist Platform for Autonomous Research* https://www.deepprinciple.com/papers/MIRA_Tech_Report.pdf.
97. Georgiev, B., Gómez-Serrano, J., Tao, T. & Wagner, A. Z. Mathematical exploration and discovery at scale. *arXiv preprint arXiv:2511.02864*. *arXiv: 2511.02864 [math.CO]* (2025).
98. Tsoukalas, G. *et al.* Advancing Mathematics Research with AI-Driven Formal Proof Search. *arXiv preprint arXiv:2605.22763*. *arXiv: 2605.22763 [cs.AI]* (2026).
99. Trinh, T. H., Wu, Y., Le, Q. V., He, H. & Luong, T. Solving olympiad geometry without human demonstrations. *Nature* **625**, 476–482. DOI: 10.1038/s41586-023-06747-5 (2024).

100. OpenAI. *Early science acceleration experiments with GPT-5* <https://cdn.openai.com/pdf/4a25f921-e4e0-479a-9b38-5367b47e8fd0/early-science-acceleration-experiments-with-gpt-5.pdf>.
101. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence* **4**, 189–191. DOI: 10.1038/s42256-022-00465-9 (2022).
102. Li, N., Pan, A., *et al.* The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. *arXiv preprint arXiv:2403.03218*. DOI: 10.48550/arXiv.2403.03218 (2024).
103. Kwa, T., West, B., Becker, J., *et al.* Measuring AI Ability to Complete Long Software Tasks. *arXiv preprint arXiv:2503.14499*. DOI: 10.48550/arXiv.2503.14499 (2025).