DO AI AGENTS HAVE RESEARCH TASTE? SENSE-BENCH FOR ANOMALY-DRIVEN SCIENTIFIC EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing AI research systems and benchmarks largely start from predefined ideas or research goals, while scientific inquiry can also begin with unexpected observations. Pursuing such observations requires research taste to judge which anomalies warrant investigation, a capability that remains insufficiently evaluated in current agents. We introduce SENSE-Bench (**S**cientific **E**xploration through **N**oticing, **S**electing, and **E**xperimenting), a benchmark for evaluating whether AI agents can recognize anomalies, judge which warrant investigation, and pursue experiments that support mechanism discovery. Each agent continues an ongoing scientific investigation from a fixed starting point. It receives records of earlier actions and observations and can run further experiments in the same environment. These environments incorporate fictional mechanisms inspired by real scientific phenomena. The benchmark also includes environments in which observations arise from random variation and those without a stable mechanism. Across these settings, known environment rules provide ground truth for evaluating agents' judgments and discoveries. The evaluation measures anomaly detection, research taste, experimental support for proposed explanations, and conclusion correctness separately. The benchmark comprises 180 cases across nine families, and the results show stronger anomaly detection and mechanism discovery in GPT-6 Astra alongside weak research taste. These findings highlight the need to improve both the selection of promising anomalies and the experiments that turn observations into supported mechanism explanations.

1 INTRODUCTION

Advances in large language models and agent technologies have enabled automated systems to conduct research and produce scientific results autonomously. These systems integrate literature search, study design, coding, experimentation, analysis, and manuscript writing (Lu et al., 2024; 2026; Ghareeb et al., 2026; Gottweis et al., 2026; Yamada et al., 2025; Yang et al., 2026; Karpathy, 2026). Despite strengths in analysis and hypothesis generation, whether they consistently match leading human researchers remains uncertain for open problem formulation, sustained exploration, and complete research delivery (Bragg et al., 2026; Xu et al., 2026; Gottweis et al., 2026). Only one of several fully generated AI Scientist manuscripts reached the average acceptance threshold in the ICLR 2025 ICBINB Workshop's initial review. Its authors judged it below the main conference's standard (Lu et al., 2026). A recent review also calls for stronger evidence of research validity, novelty, and independent verification (Ding et al., 2026). The central question is shifting from executing research steps to identifying promising leads, selecting which to pursue, and validating discoveries.

Scientific research commonly has two starting points: an idea or an anomaly (Figure 1). An idea, research goal, or theoretical hypothesis guides researchers in deriving testable predictions and designing experiments (Platt, 1964; Klahr & Dunbar, 1988). For example, a researcher may propose that one gene regulates another and test this hypothesis through interventions. An anomaly can instead prompt researchers to formulate a new question from their observations (Janz & Nylin, 2026; Kell & Oliver, 2004; Dong et al., 2026). Fleming's observation of bacterial inhibition around mold, for example, prompted investigation of the antibacterial action of penicillin (Fleming, 1929). Both starting points require judging whether a lead has research value or points to a mechanism worth

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

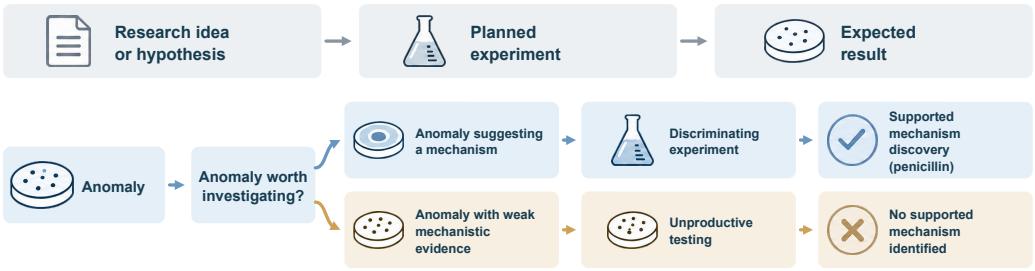


Figure 1: Two starting points for scientific inquiry. Researchers can start with a hypothesis and design experiments to test it, or encounter an anomaly and judge whether it warrants investigation. Research taste guides which anomalies to pursue, and further experiments test whether they reveal a new mechanism.

studying. For anomalies, this involves assessing the evidence for a reproducible mechanism amid random noise. We use *research taste* to describe the judgment of which anomalies deserve further investigation. Promising leads then motivate experiments to distinguish possible explanations and verify a mechanism.

Automated research follows both approaches, but most progress has come from starting with an idea. Existing systems connect supplied ideas or goals to experiments, analysis, writing, and review (Lu et al., 2024; Schmidgall et al., 2025; Tang et al., 2025; Villaescusa-Navarro et al., 2025; Yamada et al., 2025; Weng et al., 2025). Corresponding benchmarks assess task completion, experimental execution, replication, and final outputs (Chen et al., 2024; Jansen et al., 2024; Nathani et al., 2025; Huang et al., 2024; Starace et al., 2025). Automated research that starts from observations is still at an early stage. SciLink, AutoDiscovery, and OmniScientist have begun to use raw observations to formulate research questions, explore hypotheses, and guide further verification (Yao et al., 2025; Agarwal et al., 2025; Li et al., 2026a). These systems demonstrate how research can proceed from observations and evaluate the resulting discoveries or research outputs. SENSE-Bench focuses on measuring anomaly detection and selective investigation independently, and relating these abilities to subsequent mechanism discovery.

Many representative benchmarks of the complete research process overlook two key judgments: detecting an anomaly and deciding whether to investigate it. They typically supply a research question, discovery target, or interactive environment and prioritize task completion or specified outcome criteria (Chen et al., 2024; Majumder et al., 2024; Jansen et al., 2024; Bragg et al., 2026). Final outcomes can conceal process bottlenecks and different failure paths along a research trajectory (Li et al., 2026b; Wang et al., 2026). Detecting an anomaly and judging whether it warrants further resources are rarely assessed as separate capabilities. ActiveSciBench and CausalGame examine active experimentation under limited budgets, noisy observations, and mechanism recovery (Kabra et al., 2026; Chen et al., 2026). A remaining gap is to connect the two initial judgments to later experimentation and mechanism verification while allowing each stage to be assessed separately.

To address this gap, SENSE-Bench provides a benchmark for scientific discovery from anomalies across multiple domains. Each case starts from an ongoing investigation with the same prior observations for all agents, allowing fair comparison of anomaly detection. The agent continues the original task while attending to anomalies and deciding which deserve further investigation. We build interactive virtual environments for scientific research, where agents can make observations and conduct experiments. Agents must find anomalies within noisy records of interacting variables, making the search for relevant evidence part of the challenge. The environments contain fictional mechanisms inspired by real science, preserving a scientific basis for investigation while reducing reliance on answers learned during training. Agents use experiments to distinguish these mechanisms from random variation and competing explanations. We separately assess whether agents recognize anomalies (Detection), judge which warrant investigation (Taste), support their proposed mechanisms with experimental evidence (Experiment), and reach correct mechanism conclusions (Discovery).

SENSE-Bench evaluates the complete investigation, from detecting anomalies and deciding which warrant investigation to reaching mechanism conclusions through experiments. Our main contributions are:

- **A task formulation for scientific discovery from anomalies.** The task includes alternative hidden mechanisms and controls without a new mechanism, allowing similar observations to arise from mechanisms or noise. Mechanisms draw on real science but use fictional rules to reduce reliance on memorized answers and require experimental investigation. Judgments of which anomalies warrant investigation provide a concrete way to study research taste within this task.
- **A benchmark across multiple scientific domains.** The benchmark comprises nine case families in biology, medicine, earth science, chemistry, and physics. These retain the original task goals, data, and tools, with added fictional mechanisms inspired by real science, noisy observations, and opportunities for further experiments.
- **Separate assessment of research capabilities.** We assess anomaly detection, investigation decisions, and experimental support separately. Relating these assessments to final mechanism conclusions helps identify where investigations fail.

2 Related Work

Existing work develops systems that turn ideas or observations into scientific results and benchmarks that assess both research outcomes and how agents reach them.

2.1 Automated Scientific Research

Automated research systems can pursue a given research direction or develop questions from existing observations. Within a given direction, The AI Scientist and AI-Researcher automate idea generation, experimentation, and paper writing (Lu et al., 2024; Tang et al., 2025). Agent Laboratory begins with human ideas and supports human feedback during literature review, experiments, and reporting (Schmidgall et al., 2025). The AI Scientist-v2 organizes experiments through an experiment manager and progressive tree search, removing the need for human code templates (Yamada et al., 2025). Autoresearch repeatedly edits language model training code, retaining or reverting changes according to validation performance under a fixed training time budget (Karpathy, 2026). ARIS coordinates execution and review across models, checking that experimental evidence supports manuscript claims (Yang et al., 2026). Within biomedical research, this approach combines automated hypothesis development with scientists' guidance and wet experiments. AI co-scientist generates, evaluates, ranks, and refines hypotheses with scientist input, with selected hypotheses independently tested in wet laboratories (Gottweis et al., 2026). Robin combines literature search, hypothesis generation, and data analysis to investigate therapeutic candidates and their mechanisms (Ghareeb et al., 2026). Wet experiments conducted by humans provide feedback for its next research steps.

Systems that begin with observations turn patterns in data into research questions and investigate possible explanations. AutoDiscovery explores hypotheses over existing datasets through Monte Carlo tree search (Agarwal et al., 2025). Bayesian surprise guides this search by measuring how experimental results change an LLM's beliefs about the hypotheses. SciLink demonstrates a materials research workflow in three cases involving microscopy and hyperspectral spectroscopy (Yao et al., 2025). It turns features in these data into scientific claims, checks their novelty against the literature, and prepares simulation inputs or recommends further measurements. OmniScientist incorporates raw images, signals, trajectories, and other evidence throughout the research workflow (Li et al., 2026a). These observations inform the questions it asks, the experiments it conducts, and the claims in the resulting manuscripts.

2.2 Benchmarks for Scientific Research

Research evaluations assess both what agents accomplish and how they conduct their investigations. Evaluations of research outcomes cover tasks ranging from writing analysis programs and improving models to reproducing findings and recovering hidden mechanisms. Terminal-Bench-Science uses

reproducible task tests to assess analysis, simulations, proofs, code, and data artifacts across scientific workflows (Terminal-Bench-Science Team, 2026). AstaBench brings literature understanding, code execution, data analysis, and complete discovery tasks into a shared evaluation suite (Bragg et al., 2026). ResearchClawBench provides literature and raw data while withholding a target paper (Xu et al., 2026). It evaluates rediscovery of that paper's scientific findings using multimodal scoring criteria defined by experts. PaperBench supplies the target paper and uses hierarchical rubrics to assess code implementation and experimental replication (Starace et al., 2025). ScienceAgentBench asks agents to write executable programs for scientific analysis tasks drawn from published research, then checks the programs and their outputs (Chen et al., 2024). MLAgentBench and MLGym-Bench assess iterative experimentation and performance improvement in machine learning (Huang et al., 2024; Nathani et al., 2025). For interactive discovery, DiscoveryWorld provides virtual environments with scientific instruments for tasks in radioisotope dating, rocket science, and proteomics (Jansen et al., 2024). It scores task completion, relevant actions, and the explanatory knowledge agents acquire. ActiveSciBench asks agents to select experiments and recover enzyme kinetics equations or gene regulatory networks within a fixed budget (Kabra et al., 2026).

Process evaluations examine how agents choose tests and use feedback while pursuing specified inference or optimization goals. Corral provides scientific workflows and interactive tasks such as identifying unknown ions through reagent tests, determining molecular structures from spectra, and inferring hidden circuits (Ríos-García et al., 2026). Annotated trajectories and reasoning graphs reveal whether agents test candidate explanations, incorporate observations into their judgments, and revise hypotheses when evidence contradicts them. Failing to Falsify asks models to discover a numerical rule by proposing number triples and receiving feedback on whether each triple satisfies the rule (Jhaveri et al., 2026). It measures the tendency to choose tests that confirm the current hypothesis and evaluates prompting and training interventions to reduce this bias. FALSIFYBENCH asks models to discover hidden categories such as animals or foods by proposing word triples and receiving feedback on category membership (Bertolazzi et al., 2026). It examines whether models test examples outside an initially narrow hypothesis and revise their guesses after feedback. Beyond Final Scores gives agents a working but suboptimal implementation to improve under a time budget, with automated verification of successive changes (Li et al., 2026b). Its process metrics track how quickly performance improves, whether implementations run successfully, and whether agents preserve gains or recover after setbacks. Controlled comparisons further examine whether experience from earlier attempts or other tasks improves performance.

3 SENSE-BENCH

In SENSE-Bench, agents pursue an existing scientific research goal while identifying anomalies and investigating whether they reveal previously unknown mechanisms. Evaluation measures anomaly detection, judgments about which anomalies warrant investigation, experimental support for proposed explanations, and the correctness of mechanism conclusions. We adapt tasks and environments from existing benchmarks and simulation software, introducing fictional mechanisms inspired by real science. Sources include ScienceWorld (Wang et al., 2022), DiscoveryWorld (Jansen et al., 2024), ScienceAgentBench (Chen et al., 2024), Corral (Ríos-García et al., 2026), and Discover-Physics (Wiemann et al., 2026). The medical tasks build on DTRGym (Luo et al., 2024), Pulse (Webb et al., 2020), and simglucose (Xie, 2018). The benchmark comprises nine groups of cases sharing a scientific setting and environment, called *case families*, across biology, medicine, earth science, chemistry, and physics. The evaluation set contains 180 cases, with 20 cases per family. Figure 2 outlines construction, execution, and evaluation, while Appendix B summarizes the task types, sources, and experimental environments.

3.1 TASK FORMULATION

Research task. The agent takes over an ongoing scientific investigation with observations already available and opportunities to run further experiments. We refer to the agent being evaluated as the *Solver*. The investigation has an original research goal $\mathcal{G}_0$, such as studying inheritance in pea plants. An exploration instruction $\mathcal{G}_{\text{explore}}$ asks the Solver to remain curious about unexpected observations and investigate whether they indicate new mechanisms (Appendix C). Together, the research goal and exploration instruction define the full task $\mathcal{G} = \mathcal{G}_0 + \mathcal{G}_{\text{explore}}$. Here, an anomaly

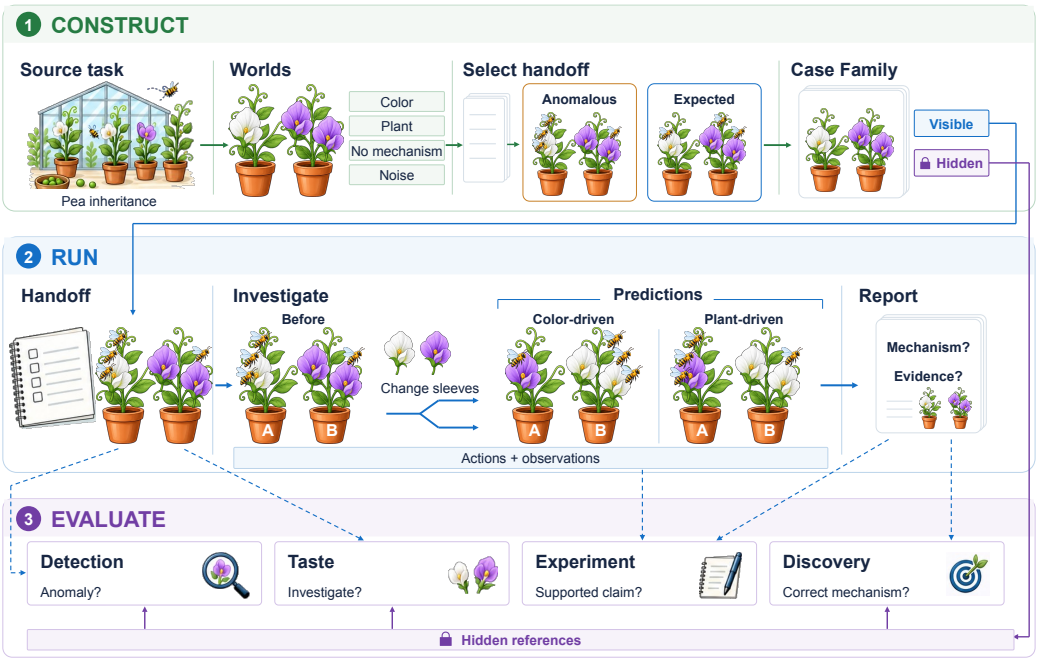


Figure 2: Construction, execution, and evaluation in SENSE-Bench. Top: mechanisms, noise, and selected handoffs define case families. Middle: the Solver conducts experiments and reports a mechanism claim. Color sleeves distinguish flower color preference from effects tied to plant or pot identity. Bottom: independent queries assess Detection and Taste at handoff and Discovery at termination. Experiment assesses support for the final claim. Connections indicate evidence sources. Hidden references and query branches remain isolated from the main investigation.

is an observation that departs from the behavior expected under the original mechanisms. Such a departure may reflect a stable new mechanism or noise, so the Solver must use further evidence to distinguish these explanations. The original goal provides the research context, while anomaly exploration is the benchmark's main evaluation target.

Handoff and interaction. Each investigation begins from the environment's initial state, where a construction agent follows the full task $\mathcal{G}$. Its actions and observations form a recorded prefix τ_{prefix}. We select a checkpoint along this trajectory and save the environment state and visible research records as the *handoff*. Formal evaluation begins when the Solver takes over from this saved state and continues under the same task instructions. The Solver receives the research instructions, prior actions and observations, existing files, current visible state, tool documentation, and remaining budget B. For each case, all Solvers start from the same evidence and experimental opportunities, allowing fair comparison of their judgments and subsequent investigations. Using the available tools, the Solver can collect observations, analyze results, run simulations, or intervene in the environment, depending on the task. Each action $a \in \mathcal{A}$ returns an observation o, forming an interaction trajectory $\tau_{\text{agent}} = (a_1, o_1, \ldots, a_n, o_n)$. The Solver chooses its next action using the evidence collected so far and continues until it stops or reaches a resource limit.

Assessment overview. We assess Detection, Taste, Experiment, and Discovery independently of progress on $\mathcal{G}_0$, using the evaluation protocol and scoring rules in Sections 3.3 and 3.4.

3.2 Benchmark Construction

We develop a pipeline that adapts source research tasks into interactive environments and creates cases with saved handoffs and reference annotations. Separate agents construct and review the materials, with human review at key stages (Appendix A).

Scenario and mechanism design. We select publicly available scientific environments that support local simulated experiments. Tasks must be complex and noisy enough that finding anomalies requires gathering, comparing, and interpreting observations. For each task, we trace how conditions produce observations and choose where to introduce a mechanism. For example, changing pollinator preferences can produce unusual flower visitation during a pea inheritance experiment. To reduce contamination from familiar mechanisms and their known consequences, hidden mechanisms are fictional but grounded in real scientific phenomena. Their rules change an interpretable process but are specific to the simulation and need not hold in reality. They cannot be inferred from background knowledge or the visible setup alone, so agents must use observations and experiments to determine whether they are present and which rules they follow. Different mechanisms may produce similar initial observations but respond differently to interventions. Agents must therefore test competing explanations. We check that existing tools distinguish the mechanisms and add tools where needed. The original goal $\mathcal{G}_0$ remains achievable in the modified environment.

Environment implementation and verification. Each run separates the adapted environment and Solver into Docker containers. Solver code runs in its own container, and experiments use documented tools without access to source code or hidden mechanism and noise settings. Independent reviewers test these access restrictions and recovery of the saved state, records, and remaining budget. Construction agents use the same tools and budget to test each mechanism. They know its configuration and must obtain observations supporting its core effect (Appendix A).

Case generation and annotation. Agents execute the complete task $\mathcal{G}$ to produce trajectories with recoverable checkpoints. Two reviewers independently identify anomalies and assess handoff viability using scientific background and possible mechanisms and noise sources, while the hidden configuration remains concealed. Arbitration resolves disagreements. Selected handoffs cover anomalous and expected observations in environments with and without a mechanism, and retain surrounding records and resources for further investigation. Annotators link each confirmed anomaly to its original records and judge whether investigation is worthwhile. Humans review all case families and anomaly annotations disputed during AI review. Final selection follows category quotas, favors diverse observations and conditions, reduces redundancy, and keeps handoff references separate from hidden mechanism truth.

3.3 Evaluation Protocol

Evaluation covers **Detection** and **Taste** at handoff, followed by **Experiment** and **Discovery** based on the completed investigation.

Assessment rationale. Anomaly status and investigation value have no perfectly sharp boundaries, and initial observations may lack the clues needed for a conclusive judgment. Our reference target is therefore the best judgment supported by the available evidence and scientific background, rather than hidden mechanism truth. The annotation procedure approximates this target without establishing a theoretical information limit.

Evaluation probes. A *run* is one Solver execution of a case. We use three isolated evaluation probes, denoted P1, P2, and P4. They ask whether the handoff record contains an anomaly, whether a confirmed anomaly warrants investigation, and what mechanism conclusion follows from the final research record. Each probe returns a structured response whose label or claim is used for scoring. The questions, response fields, and context snapshots are given in Appendix D.1. Each probe first forks the relevant handoff or final context and then asks its question in that fork. The response stays in the fork and is not returned to the main trajectory, so the probe cannot give the Solver an extra lead or information.

Constructing reference labels. During case construction, annotator agents receive an explicit background packet that the Solver does not see. It describes the scenario rules, a set of possible mechanisms that includes the embedded mechanism when one is present and extends beyond it, and plausible noise sources that extend beyond the realized noise. The packet does not identify the actual configuration. Annotators are instructed to analyze the handoff records and judge whether an observable anomaly is present and, for each confirmed anomaly, whether investigation is worthwhile.

Independent annotations aim to approach the best judgments supported by the available information. Errors and disagreements receive arbitration and human review. Final labels incorporate human corrections while retaining the original agent annotations. Taste compares the Solver's decisions with the resulting `investigate` or `do_not_investigate` labels.

Experiment assessment by an Evaluator agent. We considered discrimination through test cases and explicit predictive models for expected observations. The former requires repeated simulations across random seeds, while the latter is difficult to specify across heterogeneous scientific environments. We therefore use an independent Evaluator to assess each candidate mechanism and ordinary world alternative from the experiments visible to the Solver. It judges experimental support and discrimination among alternatives without using hidden mechanism truth.

Discovery assessment and reliability. For Discovery, an Evaluator compares the P4 conclusion with the hidden mechanism description to judge whether it identifies the defining causal features. Equivalent wording is accepted, but compatibility alone is insufficient. Annotation and Evaluator judgments remain subject to error. RQ2 examines the stability of Experiment and Discovery assessments through repeated scoring of the same saved trajectories.

3.4 METRICS

We report scores for the four assessment dimensions and two supplementary measures by family, Solver, and experimental condition. **Detection** uses the assessment returned by P1 and compares it with the construction label for observable anomaly status. `anomalous` is positive and `expected` is negative. We report F1, Cohen's κ, precision, recall, and the confusion matrix after pooling counts over runs.

Taste pairs each valid recommendation returned by P2 for a confirmed anomaly with its final triage. The positive response is `investigate`. Unresolved labels, missing responses, failures, and invalid answers are reported separately. We report F1 and Cohen's κ:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad p_o = \frac{TP + TN}{M}, \quad p_e = pq + (1-p)(1-q), \tag{1}$$

where M is the number of valid pairs, and $p = (TP+FN)/M$ and $q = (TP+FP)/M$ are the reference and Solver positive proportions, respectively. Counts are pooled across runs for each family and overall, separately by Solver and condition. With both reference classes present, always selecting or rejecting every anomaly yields $\kappa = 0$, whereas F1 can be high when every anomaly is selected.

Mechanism Selection compares the same P2 investigation decisions with whether the case contains a stable new mechanism. It reports F1 and κ on the same valid pairs as Taste, with mechanism presence as the positive reference. Taste measures agreement with investigation judgments, while Mechanism Selection measures whether investigation decisions align with actual mechanism presence.

Experiment measures how well the Solver's experiments support its final mechanism claim and distinguish it from competing explanations. It applies to stable, testable claims returned by P4, including claims outside the preselected candidates. An Evaluator uses the experimental records to judge the claim and its competing explanations. For run i, equivalent explanations are merged, leaving K_i distinct competitors, of which U_i are judged `unlikely`. If the claim is judged `plausible`, the run scores U_i/K_i, the fraction of competitors the experiments make unlikely. Otherwise, it scores 0. Competitors judged `not_assessable` remain in K_i but add no credit to U_i.

$$s_i = \mathbf{1}\{\text{target is } \texttt{plausible}\}\frac{U_i}{K_i}, \qquad \text{Experiment} = \frac{1}{N}\sum_{i=1}^{N} s_i, \tag{2}$$

where N is the number of runs with eligible claims and completed assessments. The overall score is the mean of these run scores.

Mechanism Existence measures whether the final P4 conclusion correctly judges the presence of a stable new mechanism. We report F1 with mechanism presence as the positive class, pooling confusion counts across runs, including cases without a mechanism. It complements Mechanism Selection by evaluating the existence judgment after investigation.

Discovery scores each run on cases with a stable new mechanism as 1 if the Evaluator judges that the P4 conclusion identifies its defining causal features, and 0 otherwise. Scores are averaged first across runs within each case, then across eligible cases.

See Appendix D.1 for the evaluation probes.

4 Experiments and Analysis

We first compare Solver performance across the assessment dimensions (RQ1), then examine the stability of repeated Evaluator assessments (RQ2). Finally, two archived pea investigations illustrate how experimental choices shape mechanism discovery (RQ3).

4.1 Experimental Setup

The evaluation covers 180 cases across nine case families, with 20 cases per family. We compare five Solver conditions. GPT-6 Luna and Sol use Codex at low reasoning effort, while Astra uses Codex at high reasoning effort. DeepSeek V4.1 Flash and V4 Pro use DeepSeek Harness at default settings. Each Solver performs two independent runs per case from the same frozen handoff, using the ordinary prompt and identical visible information, permissions, and budgets. The Evaluator is GPT-6 Sol with low reasoning effort, using a shared configuration across Solvers.

4.2 RQ1: Can Agents Reach Experimentally Supported Discoveries?

Table 1: Main results under the ordinary prompt with Codex for GPT and DSH for DeepSeek.

Solver	Detection		Taste		Mechanism Selection		Existence	Experiment	Discovery
	F1	κ	F1	κ	F1	κ	F1		
DeepSeek V4 Pro[†]	0.2843	0.0622	0.6915	**0.1234**	0.4247	−0.1300	0.4106	0.2076	0.0156
DeepSeek V4.1 Flash	0.4173	0.1575	0.7755	0.0422	0.5000	−0.1216	**0.4502**	0.1704	0.0417
GPT-6 Luna	0.2041	0.0696	**0.8293**	0.0029	**0.7123**	**0.0265**	0.2000	0.0000	0.0000
GPT-6 Sol	0.4569	0.1912	0.7810	0.0473	0.5098	−0.0740	0.3212	0.1291	0.0352
GPT-6 Astra	**0.5000**	**0.1959**	0.6875	−0.0938	0.3846	−0.0980	0.2105	**0.2125**	**0.2000**

[†]V4 Pro does not support image input. Scores cover eight families, excluding glacier runs with Flash.

GPT-6 Astra leads in Detection, Experiment, and Discovery (Table 1). Its Discovery score is 0.2000 versus 0.0000–0.0417 for the other reported Solvers. Its advantage in Experiment is smaller, at 0.2125 versus Pro's 0.2076. Its Detection F1 is highest at 0.5000, although its κ of 0.1959 is close to Sol's 0.1912. By contrast, Taste shows limited differentiation, with F1 scores of 0.6875–0.8293 and κ near zero across Solvers. Astra has the lowest Taste and Mechanism Selection F1 scores, indicating that stronger anomaly recognition coexists with weak initial selection. Its lower Mechanism Existence F1 also shows that judging whether a mechanism exists remains difficult.

This contrast may reflect differences in how agents use evidence gathered during subsequent investigation. Detection and Taste assess judgments at handoff, whereas discovery develops through successive experiments and observations. Despite weak initial Taste, Astra may uncover useful clues, test explanations, and revise conclusions more effectively as evidence accumulates, consistent with its Experiment and Discovery scores. We hypothesize that newer training may better support the transition from observations to mechanisms, with smaller gains in initial Taste.

4.3 RQ2: How Stable Are Repeated Evaluator Assessments?

We selected 40 saved trajectories, ten each from DeepSeek Flash, DeepSeek Pro, GPT-6 Sol, and GPT-6 Astra. GPT-6 Sol repeated scoring five times at low reasoning effort. Trajectories, probe responses, references, and prompts were fixed, while P4 interpretation, candidate preparation, and model assessments were repeated.

Scores changed little across repetitions when aggregated over the selected trajectories. Detection and Taste were identical throughout, as expected from deterministic scoring of fixed probe responses. The standard deviation of the five aggregate scores was 0.023 for Mechanism Existence F1, 0.037 for Discovery, and 0.007 for Experiment. Discovery scores were averaged over the same 24 eligible trajectories in every repetition. Experiment scores were averaged over 25 trajectories that were assessable in all five repetitions. The observed stability therefore applies to aggregate scoring on this selected sample, with greater uncertainty in individual Experiment assessments.

4.4 RQ3: Where Does Research from Anomalies Fail?

Two archived runs of GPT-6 Astra illustrate how experimental choices affect the investigation of an anomaly. In this development version of the pea environment, the original task was to determine whether tall stature was a dominant trait. In both runs, the agent started from the same records and noticed that flowers on short plants developed faster after pollination than those on tall plants. However, the short plants had white flowers and the tall plants had purple flowers, leaving height and flower color as competing explanations. The hidden mechanism accelerated development after pollination in naturally white flowers, independently of plant height.

In run A, the agent initially suspected that flowers developed faster on shorter plants. It then tested short plants with purple flowers and tall plants with white flowers, using pollen from the same donor plant. The results contradicted its prediction: purple flowers on short plants developed slowly, while white flowers on tall plants developed quickly. Repeating the comparison gave the same pattern. The agent revised its explanation and correctly identified the effect of flower color.

In run B, the agent also repeated the experiment with a common pollen donor. But it again compared short plants with white flowers against tall plants with purple flowers (Table 2). Because this comparison kept height and color linked, the agent retained the height explanation and failed to identify the correct mechanism.

Table 2: The key difference between the two investigations. Run A tested all four combinations of plant height and flower color. Run B's highlighted comparison repeated the original two combinations. Time is measured in simulation ticks from pollination until a flower formed a pod or aborted.

Plant height	Flower color	Time in run A	Time in run B
Short	White	3	3
Tall	Purple	5	5
Short	Purple	5	Not tested
Tall	White	3	Not tested

5 Discussion and Conclusion

SENSE-Bench evaluates anomaly detection, research taste, experimental support, and mechanism discovery in ongoing scientific tasks. The results show stronger Detection, Experiment, and Discovery in GPT-6 Astra alongside weak Taste, highlighting uneven performance across these capabilities. Future work will broaden the domain coverage of case families, analyze performance across mechanisms and research settings in greater detail, and explore harness designs that improve these capabilities.

AI USE STATEMENT

We use AI to develop the benchmark data pipeline and generate data, as described in the main text. It assists with translation from Chinese to English and language editing of the manuscript written by the authors.

ETHICS STATEMENT

The medical tasks use simulated environments for research evaluation. They were reviewed by a coauthor pursuing a PhD in a clinical discipline.

REPRODUCIBILITY STATEMENT

The task, evaluation protocol, and metrics are defined in Sections 3.1, 3.3, and 3.4. Construction details, family descriptions, and exploration prompts appear in Appendices A, B, and C. Evaluation probe inputs and questions are given in Appendix D.1.

REFERENCES

Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, and Peter Clark. Autodiscovery: Open-ended scientific discovery via bayesian surprise. In *Advances in Neural Information Processing Systems*, volume 38, pp. 28504–28542, 2025. doi: 10.52202/085713-0847. URL https://papers.neurips.cc/paper_files/paper/2025/hash/23b127521af7ca7a42f5cdb7507be4f2-Abstract-Conference.html.

Leonardo Bertolazzi, Katya Tentori, and Raffaella Bernardi. FALSIFYBENCH: Evaluating inductive reasoning in LLMs with rule discovery games, 2026. URL https://arxiv.org/abs/2606.04751.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefan Candra, Jason Dunkelberger, Dan Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Cecile Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel S. Weld. Astabench: Rigorous benchmarking of AI agents with a scientific research suite. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=M7TNf5J26u.

Zhenhao Chen, Yongqiang Chen, Chenxi Liu, Junchi Yu, Xiangchen Song, Zijian Li, Jialin Li, Philip Torr, Bo Han, and Kun Zhang. Causalgame: Benchmarking causal thinking of LLM agents in games. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2607.04293.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery, 2024. URL https://arxiv.org/abs/2410.05080.

Tianyu Ding, Aditya Nannapaneni, Bingfan Liu, and Ling Zhang. Autonomous research agents: A survey of AI scientists and the verification gap, 2026. URL https://arxiv.org/abs/2608.05179.

Yiru Dong, Richong Zhang, Fanshuang Kong, and Si Chen. ExeCRE: Execution-consistency guided reliability estimation for self-correcting code generation. In *Proceedings of the 41st IEEE/ACM International Conference on Automated Software Engineering*, ASE '26, Munich,

Germany, October 2026. Association for Computing Machinery. ISBN 979-8-4007-2882-2. doi: 10.1145/3832783.3837445. URL https://doi.org/10.1145/3832783.3837445.

Alexander Fleming. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of b. influenzae, 1929. URL https://europepmc.org/article/MED/11545337.

Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodriques. A multi-agent system for automating scientific discovery. *Nature*, 655:497–505, 2026. doi: 10.1038/s41586-026-10652-y. URL https://doi.org/10.1038/s41586-026-10652-y.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with co-scientist. *Nature*, 655(8122):487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL https://doi.org/10.1038/s41586-026-10644-y.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: Evaluating language agents on machine learning experimentation, 2024. URL https://proceedings.mlr.press/v235/huang24y.html.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents, 2024. URL https://arxiv.org/abs/2406.06769.

Niklas Janz and Sören Nylin. Observation-driven science. In *Understanding Science: The Web of Trust*, pp. 12–27. Cambridge University Press, 2026. doi: 10.1017/9781009672863.003. URL https://doi.org/10.1017/9781009672863.003.

Ayush Rajesh Jhaveri, Anthony GX-Chen, Ilia Sucholutsky, and Eunsol Choi. Failing to falsify: Evaluating and mitigating confirmation bias in language models, 2026. URL https://arxiv.org/abs/2604.02485.

Sanchit Kabra, Nikhil Abhyankar, Saaketh Desai, Prasad Iyer, and Chandan K Reddy. LLM-autoscilab: Closed-loop scientific discovery via active experimentation with LLMs, 2026. URL https://arxiv.org/abs/2605.24043.

Andrej Karpathy. autoresearch: AI agents running research on single-GPU nanochat training automatically. GitHub repository, 2026. URL https://github.com/karpathy/autoresearch.

Douglas B. Kell and Stephen G. Oliver. Here is the evidence, now what is the hypothesis? the complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*, 26(1):99–105, 2004. doi: 10.1002/bies.10385. URL https://doi.org/10.1002/bies.10385.

David Klahr and Kevin Dunbar. Dual space search during scientific reasoning, 1988. URL https://doi.org/10.1207/s15516709cog1201_1.

Bobo Li, Hao Fei, Tianjie Ju, Mong-Li Lee, and Wynne Hsu. OmniScientist: An omni-modal omni-discipline AI scientist, 2026a. URL https://arxiv.org/abs/2608.13558.

Yiwei Li, Wanli Yang, Hexiang Tan, Xiangzhou Huang, Zhengyu Chen, Ziran Li, Borun Chen, Shanglin Lei, Huaisheng Zhu, Hao Tian, Fei Sun, Xunliang Cai, and Jingang Wang. Beyond final scores: A systematic evaluation of agents for long-horizon AI research and development, 2026b. URL https://arxiv.org/abs/2608.13417.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery, 2024. URL https://arxiv.org/abs/2408.06292.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651:914–919, 2026. doi: 10.1038/s41586-026-10265-5. URL https://doi.org/10.1038/s41586-026-10265-5.

Zhiyao Luo, Mingcheng Zhu, Fenglin Liu, Jiali Li, Yangchen Pan, Jiandong Zhou, and Tingting Zhu. DTR-Bench: An in silico environment and benchmark platform for reinforcement learning based dynamic treatment regime. *arXiv preprint arXiv:2405.18610*, 2024. URL https://arxiv.org/abs/2405.18610.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models, 2024. URL https://arxiv.org/abs/2407.01725.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. MLGym: A new framework and benchmark for advancing AI research agents, 2025. URL https://arxiv.org/abs/2502.14499.

John R. Platt. Strong inference, 1964. URL https://doi.org/10.1126/science.146.3642.347.

Martiño Ríos-García, Nawaf Alampara, Chandan Gupta, Indrajeet Mandal, Sajid Mannan, Ali Asghar Aghajani, N. M. Anoop Krishnan, and Kevin Maik Jablonka. AI scientists produce results without reasoning scientifically, 2026. URL https://arxiv.org/abs/2604.18805.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043. Association for Computational Linguistics, November 2025. doi: 10.18653/v1/2025.findings-emnlp.320. URL https://aclanthology.org/2025.findings-emnlp.320/.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. Paperbench: Evaluating AI's ability to replicate AI research, 2025. URL https://arxiv.org/abs/2504.01848.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. AI-researcher: Autonomous scientific innovation, 2025. URL https://arxiv.org/abs/2505.18705.

Terminal-Bench-Science Team. Terminal-Bench-Science: Evaluating AI agents on research workflows across scientific domains, August 2026. URL https://github.com/harbor-framework/terminal-bench-science.

Francisco Villaescusa-Navarro, Boris Bolliet, Pablo Villanueva-Domingo, Adrian E. Bayer, Aidan Acquah, Chetana Amancharla, Almog Barzilay-Siegal, Pablo Bermejo, Camille Bilodeau, Pablo Cárdenas Ramírez, Miles Cranmer, Urbano L. França, ChangHoon Hahn, Yan-Fei Jiang, Raul Jimenez, Jun-Young Lee, Antonio Lerario, Osman Mamun, Thomas Meier, Anupam A. Ojha, Pavlos Protopapas, Shimanto Roy, David N. Spergel, Pedro Tarancón-Álvarez, Ujjwal Tiwari, Matteo Viel, Digvijay Wadekar, Chi Wang, Bonny Y. Wang, Licong Xu, Yossi Yovel,

Shuwen Yue, Wen-Han Zhou, Qiyao Zhu, Jiajun Zou, and Íñigo Zubeldia. The denario project: Deep knowledge AI agents for scientific discovery, 2025. URL https://arxiv.org/abs/2510.26887.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, Eric P. Xing, and Zhiting Hu. FIRE-bench: Evaluating agents on the rediscovery of scientific insights, 2026. URL https://arxiv.org/abs/2602.02905.

Jeffrey B. Webb, Aaron Bray, Philip K. Asare, Rachel B. Clipp, Yatin B. Mehta, Sudheer Penupolu, Aalpen A. Patel, and S. Mark Poler. Computational simulation to assess patient safety of uncompensated COVID-19 two-patient ventilator sharing using the Pulse Physiology Engine. *PLOS ONE*, 15(11):e0242532, 2020. doi: 10.1371/journal.pone.0242532. URL https://doi.org/10.1371/journal.pone.0242532.

Yixuan Weng, Minjun Zhu, Qiujie Xie, Qiyao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. Deepscientist: Advancing frontier-pushing scientific findings progressively, 2025. URL https://arxiv.org/abs/2509.26603.

Matt L. Wiemann, Lindsay M. Smith, Peter Melchior, Siddharth Mishra-Sharma, Andrew Gordon Wilson, Pavel Izmailov, and Carolina Cuesta-Lázaro. Discoverphysics: Benchmarking LLMs for out-of-the-box scientific thinking, 2026. URL https://arxiv.org/abs/2605.26087.

Jinyu Xie. Simglucose v0.2.1, 2018. URL https://github.com/jxx123/simglucose. Accessed 2026-09-25.

Wanghan Xu, Shuo Li, Tianlin Ye, Qinglong Cao, Yixin Chen, Hengjian Gao, Yiheng Wang, Qi Li, Kun Li, Sheng Xu, Shengdu Chai, Fangchen Yu, Xiangyu Zhao, Zhangrui Zhao, Weijie Ma, Zijie Guo, Koutian Wu, Haoyu Zhou, Haoxiang Yin, Lixue Cheng, Chaofan Hu, Haoxuan Li, Lu Mi, Xuxuan Xie, Yifan Zhou, Ruizhe Chen, Zhiwang Zhou, Xingjian Guo, Yuhao Zhou, Xuming He, Shengyuan Xu, Xinyu Gu, Jiamin Wu, Mianxin Liu, Chunfeng Song, Fenghua Ling, Dongzhan Zhou, Shixiang Tang, Yuqiang Li, Mao Su, Peng Ye, Siqi Sun, Bin Wang, Xue Yang, Zhenfei Yin, Tianfan Fu, Guangtao Zhai, Wanli Ouyang, Bo Zhang, Lei Bai, and Wenlong Zhang. Researchclawbench: A benchmark for end-to-end autonomous scientific research, 2026. URL https://arxiv.org/abs/2606.07591.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search, 2025. URL https://arxiv.org/abs/2504.08066.

Ruofeng Yang, Yongcan Li, and Shuai Li. ARIS: Autonomous research via adversarial multi-agent collaboration, 2026. URL https://arxiv.org/abs/2605.03042.

Lance Yao, Suman Samantray, Ayana Ghosh, Kevin Roccapriore, Libor Kovarik, Sarah Allec, and Maxim Ziatdinov. Operationalizing serendipity: Multi-agent AI workflows for enhanced materials characterization with theory-in-the-loop, 2025. URL https://arxiv.org/abs/2508.06569.

Leijun Zhou, Zhihao Liu, Xiang Qu, Chenxu Liu, Yifei Liu, Yanke Yu, Jingzhe Xu, Xuejun Wu, Buyue Qian, Xi Chen, Yaowei Zheng, and Junhao Hu. GDPevo: Evaluating agent self-evolution on real business tasks. *arXiv preprint arXiv:2608.03764*, 2026. URL https://arxiv.org/abs/2608.03764.

A BENCHMARK CONSTRUCTION

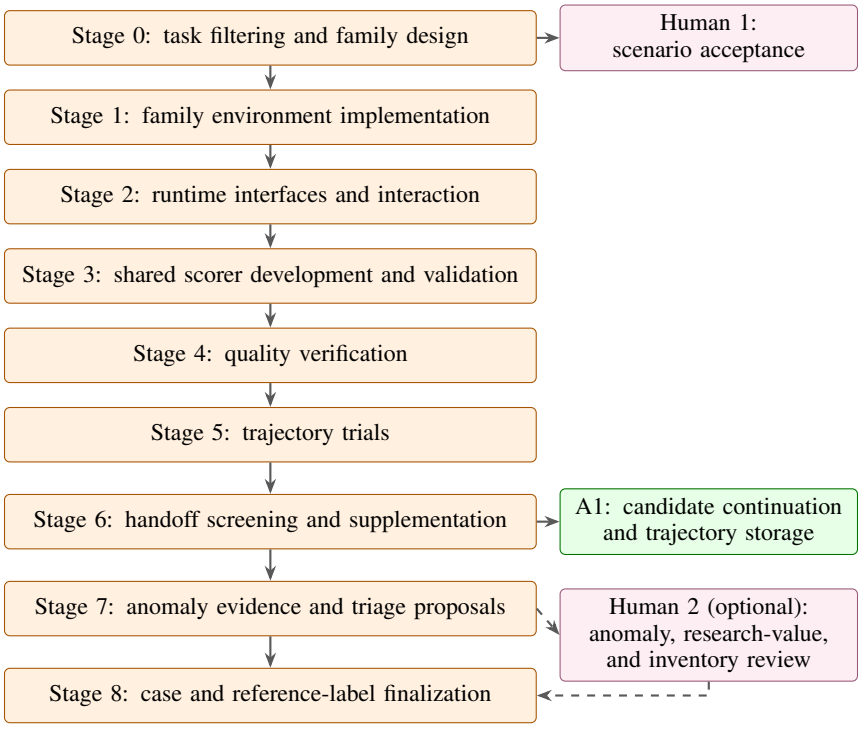


Figure 3: Construction workflow. Dashed links indicate optional human review. A1 stores candidate continuation trajectories.

Figure 3 summarizes construction from task selection to frozen cases and reference labels. Separate agents build and review environments, following an approach inspired by GDPevo (Zhou et al., 2026). Verification checks implementation, information isolation, and state recovery. Designers test each mechanism with the Solver's tools and budget before candidate handoffs are screened and annotated. Human review covers all families and disputed anomaly labels. Formal evaluation uses the finalized cases and labels.

B CASE FAMILIES

Table 3: Sources, task types, and experimental environments of the nine case families. We adapt source tasks and extend their environments with hidden mechanisms.

Domain	Task	Source	Experimental environment
Biology	Pea inheritance	ScienceWorld (Wang et al., 2022)	Grow and pollinate plants, manipulate flower appearance, and observe visits and pod formation.
Biology	Food and health	DiscoveryWorld (Jansen et al., 2024)	Test mushroom samples and track illness through instruments, treatments, and eating events.
Medicine	Chemotherapy	DTRGym (Luo et al., 2024)	Set drug doses and observe tumor, normal and immune cells, and drug concentration.
Medicine	Shared ventilation	Pulse (Webb et al., 2020)	Adjust a shared ventilator and compare two simulated patients' ventilation and gas exchange.
Medicine	Glucose control	simglucose (Xie, 2018)	Deliver insulin to virtual patients and monitor glucose trajectories and hypoglycemia risk.
Earth science	Glacier evolution	ScienceAgentBench (Chen et al., 2024)	Simulate glacier area and thickness under climate forcing and compare spatial states over time.
Chemistry	Pure salt identification	Corral (Ríos-García et al., 2026)	Identify an unknown salt with reagents, precipitation, acid–base tests, and flame tests.
Chemistry	Mixed cation identification	Corral (Ríos-García et al., 2026)	Identify mixed ions through aliquoting, precipitation, filtration, complexation, and flame tests.
Physics	Particle probes	DiscoverPhysics (Wiemann et al., 2026)	Place probes in a particle simulation and infer motion rules from their trajectories.

C EXPLORATION PROMPT

All case families use the following ordinary exploration instruction, with identical wording before and after handoff.

> While completing the task, remain curious about unexpected findings and investigate whether they point to new mechanisms.

D ADDITIONAL EVALUATION DETAILS

D.1 CONTEXT SNAPSHOTS AND EVALUATION PROBES

Each probe runs in an independent copy of the relevant context using the same Solver as the main investigation. Snapshots preserve visible instructions, materials, interaction history, and restoration information, but exclude hidden truth. Probe responses remain within their branches, so evaluation does not guide the main investigation. P1 uses the handoff context once per run. P2 uses a separate copy for each confirmed anomaly, identified by `case_id` and `anomaly_id`. P4 uses the final context and returns a structured judgment and a free-text `conclusion`, which Experiment and Discovery assess.

P2 supplies original records that jointly contain the anomaly, preserving conditions, controls, and outcomes. Localization uses complete interaction turns, or a complete tool response when a turn is too long. Program-assigned identifiers retain the original context. No anomaly description, variable hint, derived comparison, or reference answer is supplied. The executed probe questions follow. The selected records are appended to P2. Exploration instructions appear separately in Appendix C.

P1.

Using only the information currently available, determine whether the results are consistent with expectations or anomalous. Briefly state the reason for your assessment, then return only the following JSON object. Set `assessment` to either `expected` or `anomalous`.

P2.

The following group of original observations jointly contains an anomaly. Using these records and the available context, decide whether further experimental investigation is worthwhile. Briefly state the reason for your recommendation, then return only the specified JSON object. Set `recommendation` to either `investigate` or `do_not_investigate`.

P4.

Based on the available research record, determine whether the results reveal a new mechanism or are better explained by known behavior, noise, or random variation. Briefly state the reason for your conclusion. If a new mechanism has been identified, describe it; otherwise, give the most likely explanation. Return only the following JSON object. Set `conclusion_type` to either `new_mechanism` or `no_new_mechanism`.