DO LLMS SEE THE GORILLA IN THE DATA? A BENCHMARK FOR OFF-OBJECTIVE SCIENTIFIC DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific discovery depends on achieving stated objectives and also on noticing what those objectives leave unstated. We call this ability off-objective discovery, the recognition, while pursuing a given research goal, of an important finding that the same materials support but the goal never mentions. Such discoveries are largely attributed to human serendipity, and existing benchmarks evaluate only whether agents complete the objective they were given. We introduce UncoverBench, a benchmark that measures this ability against verifiable answers. Each of its 193 tasks is built from a recent peer-reviewed paper. A model receives the paper's tables as the working data of an ongoing study, together with a research objective, and is asked what should be done next. One of the paper's central findings is never mentioned, yet the tables support it. Every task also carries decoy conclusions that the tables contradict, which catch reports that assert patterns without checking them. Across twelve frontier models, reports complete the stated objective about three times in four but state the unstated finding in only 27.9% of cases. The misses are failures of recognition. In a sample labelled by hand, about two thirds of reports interpret the evidence for the finding, yet fewer than one in four state it, and models that fetch the right tables on their own still miss the finding in about half of the tasks or more. A generic reminder to look beyond the objective raises discovery to 43.5% and leaves decoy endorsement unchanged. Off-objective discovery thus remains a distinct capability that current frontier models largely lack, whether they read all materials at once or fetch the evidence with tools. UncoverBench provides a verifiable foundation for measuring and improving this capability, moving research agents from completing the objectives they are given toward noticing the findings those objectives leave unstated.

1 INTRODUCTION

Scientific progress comes both from answering the questions that researchers set out to answer and from noticing what those questions leave out (Yaqub, 2018). We call the second capacity off-objective discovery, the recognition, while pursuing a given research objective, of an important finding that the same materials support but the objective never mentions. Recent progress in autonomous research systems has largely focused on carrying a given objective through to a validated result, in machine learning research (Lu et al., 2026; Tang et al., 2026), biomedical hypothesis generation (Gottweis et al., 2026; Ghareeb et al., 2026), large-scale data analysis (Mitchener et al., 2025b), and the improvement of scientific software, algorithms and mathematical results (Aygün et al., 2026; Novikov et al., 2025; Sothanaphan, 2026). Their evaluations follow the same design. Benchmarks score whether a system completes an engineering task, improves a metric or reproduces a paper (Chan et al., 2025; Wijk et al., 2025; Starace et al., 2025), recovers the result of a hidden target study (Wang et al., 2026c), or answers a research question posed over data (Majumder et al., 2025; Chen et al., 2025; Panigrahi et al., 2026), and in every case the finding to be recovered is the one that the task asks for. Off-objective discovery remains unmeasured.

The most exciting thing in science is not "Eureka" but "That's funny". Discovery begins with an act of recognition, and recognition is not a matter of luck. Accounts of serendipity in science find that

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.

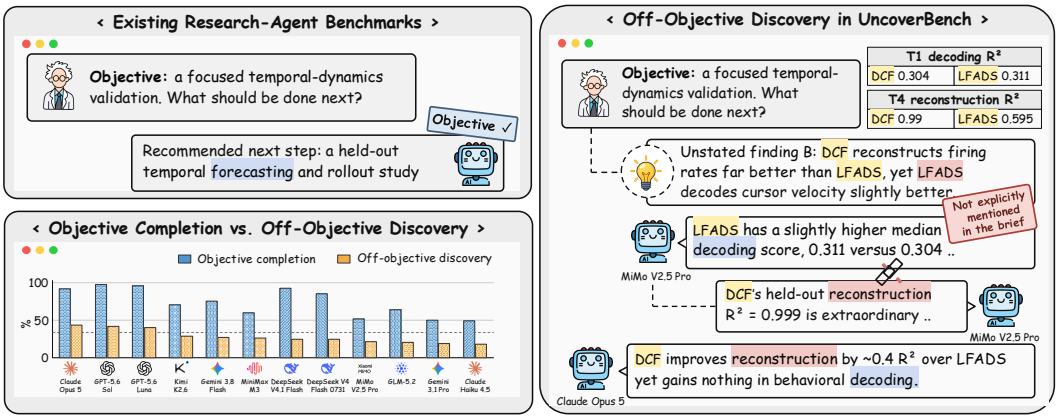

Figure 1: Existing benchmarks and UncoverBench score the same task. Existing research-agent benchmarks score whether a report completes the objective it was given (top left). In UncoverBench, one of the source paper's central findings is supported by the tables yet never mentioned in the brief (right). Both reports, written with the reminder, complete the objective. MiMo V2.5 Pro reads the decoding and reconstruction results and singles out the near-perfect reconstruction for further study without setting it against decoding, and Claude Opus 5 states the finding. Across the twelve models, most reports complete the objective, and without the reminder no model states the unstated finding in half of its reports (bottom left, the numbers of Table 2).

an unexpected observation becomes a discovery only when a researcher recognises its significance, a pattern that recurs across fields (Yaqub, 2018). Recognition is also what a focused objective suppresses. Observers busy with a demanding task miss an unexpected event in plain view (Simons & Chabris, 1999), and students who analysed a dataset with a specific hypothesis in mind were far less likely to notice a gorilla drawn into its scatter plot than students who explored the same data freely (Yanai & Lercher, 2020). Measuring this capacity requires knowing in advance which important findings the materials hold, because a request for anything interesting in the data rewards volume with answers that cannot be checked. Published science supplies such answer keys.

We introduce UncoverBench, which turns a peer-reviewed paper into an off-objective discovery task by withholding one of its central claims and giving the model an objective that never mentions it. Each task presents the tables of a recent paper as the working data of an ongoing study, together with an objective that the study would naturally pursue next, and asks the model only what should be done next (Figure 1). The withheld claim is absent from the brief, yet the tables support it. Its importance is therefore set by the authors and the reviewers who accepted the paper, and every report can be checked against a fixed statement. Every task also carries three decoys, conclusions that look supported by the tables but that the tables contradict, so that a report that asserts patterns without checking them is caught. The brief is written by a model that never sees the finding, and every task is reviewed by hand.

We evaluate twelve frontier models, six proprietary and six with open weights. Our results show that off-objective discovery remains difficult for every evaluated model. Reports complete the stated objective about three times in four, yet they state the unstated finding in 27.9% of cases, and no model reaches one half. To locate where discovery fails, we place human-labelled reports at four levels, from never touching the evidence to stating the finding. We find that the misses are failures of recognition. About two thirds of these reports interpret the evidence for the finding, while fewer than one in four state it. The gap persists when models choose which tables to read, when the evidence sits in the very tables that the objective names, and however long models reason, and it widens as the materials grow. A generic reminder to look beyond the objective raises discovery to 43.5% with decoy endorsement unchanged, yet reports that interpret the evidence without stating the finding remain the largest group. Taken together, these findings show that noticing what lies beyond an objective is a capability distinct from completing it, and one that current evaluations of research agents leave unmeasured.

Our contributions are summarized as follows.

Table 1: UncoverBench among evaluations of research agents. Columns are benchmark families with representative members. ✓ holds for the family, ○ for some members, ✗ not at all.

	Objective completion MLE-bench, RE-Bench, MLGym	Target-study rediscovery PaperBench, FIRE-Bench	Data-driven insight DiscoveryBench, InsightBench, HeurekaBench	UncoverBench (ours)
Built on real research tasks or data	✓	✓	✓	✓
Ground truth fixed by peer-reviewed findings	✗	✓	○	✓
Scored finding lies outside the stated objective	✗	✗	✗	✓
Objective and discovery scored separately	✗	✗	✗	✓

- **Off-objective discovery as an evaluation target.** We formulate off-objective discovery as recovering a withheld central claim of a published paper while pursuing an objective that never mentions it, and separate it from completing the objective and from rediscovering the finding that a task asks for (Table 1).
- **UncoverBench.** We build 193 tasks from recent peer-reviewed papers, each with a brief written without access to the finding and three decoys that the tables contradict, and score reports with two large language model (LLM) judges whose verdicts agree with blind human annotation about as closely as two annotators agree with each other.
- **Empirical study.** We evaluate twelve frontier models and locate their failure in recognition. Models read the evidence yet do not state the finding, and the gap holds across evidence access, recall and reasoning length.

2 Related Work

Autonomous research and its evaluation. Systems that automate research now span the full cycle from idea to paper (Lu et al., 2024; Yamada et al., 2025; Lu et al., 2026; Schmidgall et al., 2025; Weng et al., 2026), generate and test biomedical hypotheses (Gottweis et al., 2026; Ghareeb et al., 2026; Swanson et al., 2025; Mitchener et al., 2025b), and search over programs to improve scientific software and algorithms (Novikov et al., 2025; Aygün et al., 2026). Their evaluations follow the objective. Engineering benchmarks score whether an agent completes a task or improves a metric (Huang et al., 2024; Chan et al., 2025; Wijk et al., 2025; Nathani et al., 2025; Garikaparthi et al., 2026), replication benchmarks score whether it reproduces a given paper (Starace et al., 2025; Siegel et al., 2024), suites such as AstaBench cover the stages of scientific research assistance (Bragg et al., 2026), and end-to-end studies ask reviewers or LLMs to judge the papers an agent writes (Bianchi et al., 2026; Tang et al., 2026). Recent analyses find that agents complete the engineering of a research project yet make little progress on its open questions (Kirgis et al., 2026), and that they elaborate locally more than they broaden exploration (Tang & Yang, 2026). FIRE-Bench also takes its ground truth from published findings and asks agents to rediscover the finding that its research question points to (Wang et al., 2026c). UncoverBench places that finding outside the question (Table 1).

Data-driven discovery and insight. A second line tests whether models draw findings from data. DiscoveryBench, BLADE and BixBench pose research questions over datasets drawn from the literature (Majumder et al., 2025; Gu et al., 2024; Mitchener et al., 2025a), ScienceAgentBench asks for analysis programs on tasks extracted from published studies (Chen et al., 2025), InsightBench plants insights in business analytics data (Sahu et al., 2025), and HeurekaBench grounds open-ended questions in single-cell studies and their code (Panigrahi et al., 2026). NewtonBench and HypoBench ask whether models recover laws or hypotheses from observations (Zheng et al., 2026; Liu et al., 2025), and DiscoveryWorld places agents in simulated discovery tasks (Jansen et al., 2024). Work on ideation asks whether models propose or judge research ideas that experts and the community value (Si et al., 2025; 2026; Tong et al., 2026). TruthInsightBench exposes only a neutral objective and frozen data and scores whatever claim the agent decides the data support (Yang et al., 2026). In all of these settings the scored finding is the one the task asks for or the one the agent chooses. UncoverBench asks whether a model working on one objective recognises a finding that the objective does not mention.

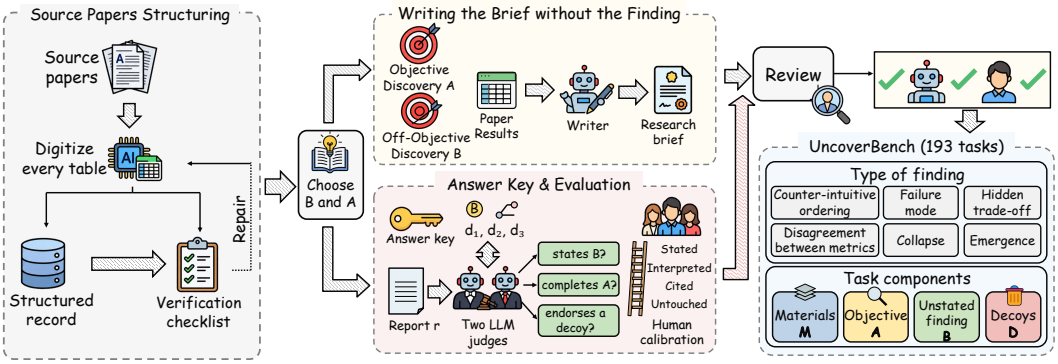

Figure 2: Construction and scoring of UncoverBench. A paper's central claim becomes the unstated finding B, a writer that never sees B produces the brief, and two LLM judges score each report.

Noticing beyond the instruction. Psychology has long shown that a focused task hides unexpected events from the observer (Simons & Chabris, 1999) and that a hypothesis can hide a pattern in data from the analyst who holds it (Yanai & Lercher, 2020). Studies of serendipity describe how discoveries arise when researchers notice what they were not looking for (Yaqub, 2018). For language models, recent benchmarks test whether they pick up implicit cues under explicit instructions (Cai et al., 2026), question flawed premises they were not asked about (Li et al., 2026), catch details that experienced researchers would catch (Wang et al., 2026a), and act on needs that a user left unstated (Lu et al., 2025; Pasternak et al., 2025; Harfi et al., 2026). Other systems pursue surprise on purpose, by maximising Bayesian surprise over hypotheses (Agarwal et al., 2025) or by answering questions designed to reveal unexpected connections (Wang et al., 2026b), which makes surprise itself the objective. UncoverBench combines a stated research objective with an unstated finding whose truth and importance are fixed in advance by the source paper.

3 UncoverBench: From Papers to Off-Objective Discovery Tasks

UncoverBench turns published papers into research tasks in which an important finding is present in the materials and absent from the objective.

3.1 Task Formulation

Off-objective discovery involves three things that ordinary evaluations treat as one, the materials a researcher works with, the objective the researcher pursues, and the findings that the materials support. We keep the three apart.

A task is a tuple (M, A, B, D). The materials M describe an ongoing study and hold its experimental records, here the numerical tables of a paper together with the settings needed to read them. The objective A states what the study is currently trying to achieve. The unstated finding B is a claim with three properties. It is supported by M, so a careful reader of the tables can reach it. It lies outside A, so the objective can be completed without it. Writing $\mathcal{R}_A$ for the set of reports that complete A, some $r \in \mathcal{R}_A$ does not state B. It is important, which we make concrete by requiring B to be one of the central claims that the source paper makes about its own results. The decoys $D = \{d_1, d_2, d_3\}$ are claims that look supported by M but are contradicted by it.

The model receives M and A with a single fixed request that asks what should be done next and why, and it writes a free-form report r. We score r on three quantities. Objective completion records whether r reaches a sound conclusion about A. Discovery records whether r itself states B. Decoy endorsement records whether r asserts any d_i as an established result. An ideal report completes A and states B. The finding need not become the main conclusion, and the task never asks for it. Discovery is scored on what the report states, because the report is all that an autonomous system returns, and a finding it does not report reaches no one.

This formulation separates off-objective discovery from ordinary task performance, which objective completion measures, and from rediscovery, in which the finding to be recovered is the one the task asks for (Wang et al., 2026c). Objective completion and decoy endorsement together confirm that a model reached B without abandoning its task and without asserting claims that the tables contradict.

3.2 BENCHMARK CONSTRUCTION

Each task is built in five steps (Figure 2), and every step is arranged so that the finding enters the task only through the numbers.

Source papers. Papers come from recent machine learning venues, mostly the main track of ICML 2026 and also ICML 2025, ICLR 2026, NeurIPS 2025 and ICCV 2025. Their results sit in dense numerical tables that can be checked cell by cell. Peer review gives the ground truth two further properties. Its central claims have been examined by expert reviewers, and they are the claims that the authors themselves chose to put forward.

Digitising every table. An LLM (DeepSeek V4 Flash) converts each paper into a structured record that holds every table in the main text and the appendix, with each cell copied exactly as printed. The record also lists the claims that the paper makes about its own results, each with a verbatim quote from the abstract, the contribution list, the introduction or the conclusion. No table is selected at this stage, so the record does not favour any finding. A program then checks that every numeric cell appears in the paper text, that every table the paper refers to is accounted for, and that every quote is verbatim.

Choosing the finding and the objective. A second model (GPT-5.6 Sol) selects as B one of the paper's claims that the printed numbers support. It must keep B faithful to the paper, without caveats, exceptions or reinterpretations of its own, and it may not use an overall ranking of methods as B. The same step proposes the objective A as the natural continuation of the study, such as the next training iteration or the next round of analysis. A must be something a researcher could advance without stating B, and it may not reduce to picking a winner among the listed methods.

Writing the brief without the finding. The public brief is written by a model that never sees B or the answer key. It receives the objective and the tables with their captions and notes, and from them it writes the research context and the definitions a reader needs to interpret the settings. It summarises no result. The complete numerical tables are then appended in the paper's original order, with no rows or columns chosen for B, and the brief carries no paper title or author names. Every brief ends with the same request for next steps, and none mentions further findings, anomalies or B. The finding is present only as a relation among cells, often across tables that the objective never names.

Decoys. For each task the construction model also writes three decoys, misreadings that a hurried analyst could draw from the tables and that the tables contradict, each with the reason it is false and a rule for when a report endorses it.

Review. Every task is reviewed by hand. Before any model is evaluated, an LLM reviewer screens each task against the requirements above and a human reviewer then examines it again.

3.3 EVALUATION PROTOCOL

The discovery criterion. A report discovers B when it states the relation that B describes, with the right objects and the right direction. A statement made in passing, as part of the argument for A, or with a hedge all count. The report need not call it a finding, list every supporting number, explain it or propose follow-up work. A report that mentions the two halves of B in different places and leaves the reader to connect them does not count. This rule separates recognising a finding from reading the numbers that contain it.

To locate where discovery fails in a sample of reports, annotators place each report at one of four levels. A STATED report states B, an INTERPRETED report reads and interprets the evidence for B without stating it, a CITED report only quotes the relevant numbers, and an UNTOUCHED report

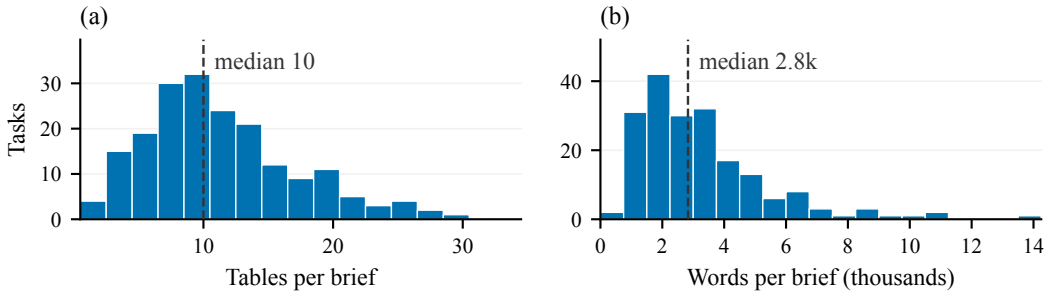

Figure 3: Size of the materials across the 193 tasks. (a) Number of tables per brief. (b) Length of the brief. Dashed lines mark the medians.

does not refer to the evidence at all. Only STATED counts as discovery. The other levels show how far a model got before it stopped.

Judges and human validation. Two LLM judges (Zheng et al., 2023), GPT-5.6 Sol and Claude Opus 5, read every full report together with the answer key and decide whether it states B, whether it completes A and whether it endorses a decoy. A report counts as stating B when both judges credit it, and as completing A when both judges grade it as complete. Neither judge can thus credit a report alone, even one from its own model family. Two annotators validated this rule on 200 reports, labelling each independently without seeing the model identity, the condition or the judges' verdicts. The reports come from 50 tasks drawn at random, each with two models under both conditions, and cover all twelve models. On whether a report states B, the judges agree with the two annotators at Cohen's κ of 0.85 and 0.83 (Cohen, 1960), close to the 0.87 between the annotators themselves, with a precision of 0.92 and 0.90 and a recall of 0.87 for both. The agreement is similar across model families and across the two conditions, with κ between 0.81 and 0.86 (Appendix B). On objective completion the judges pass 73.5% of the labelled reports and the annotators 74.5% and 74.0%.

Objective completion and decoy endorsement. A report completes the objective when it reaches a clear conclusion consistent with the reference continuation or an equivalent one. A report endorses a decoy only when it asserts the decoy as an established result, not when it mentions, tests or rejects it. We count a report as endorsing a decoy when either judge flags it, which gives an upper bound.

3.4 BENCHMARK STATISTICS AND RELIABILITY

UncoverBench contains 193 tasks from 193 distinct papers, 175 of them from ICML 2026. A brief holds 11.4 tables on average (median 10, range 1 to 35) and 3,284 words (median 2,830), with a long tail in both (Figure 3). An LLM classifier labels each finding with its type, the largest being counter-intuitive orderings (88 tasks), and with where its evidence sits (Table 6). In 124 tasks the evidence for B lies entirely in tables that the objective does not name, in 31 it lies partly there, and in 38 it lies entirely in tables that the objective names.

The two annotators agree on whether a report states B for 94.5% of reports ($\kappa = 0.87$) and reach a quadratically weighted κ of 0.97 on the four levels (Cohen, 1968). On whether a report completes the objective they agree on 97.5% of reports ($\kappa = 0.93$).

4 EXPERIMENTAL SETUP

Models. We evaluate twelve frontier models, grouped as in Table 2. The proprietary models are GPT-5.6 Sol, GPT-5.6 Luna, Claude Opus 5, Claude Haiku 4.5, Gemini 3.8 Flash and Gemini 3.1 Pro. The open-weight models are GLM-5.2, MiniMax M3, DeepSeek V4.1 Flash, DeepSeek V4 Flash 0731, Kimi K2.6 and MiMo V2.5 Pro. All models are queried in a single turn through one application programming interface (API) gateway. The GPT-5.6 models run at temperature 1.0 and the others at their provider defaults, with an output budget of 32,768 tokens (Appendix A). Each model answers every task once in each condition, which gives 4,632 reports.

Table 2: Main results (%). Objective is the share of reports that complete the stated objective, and discovery is the share that state the unstated finding, both as scored by the two judges (Section 3.3). The middle six columns give discovery by type of finding without the reminder. Bold marks the best value in each group.

Models	Objective	Discovery by Finding Type (No Reminder)						Overall Discovery		
		Counter-intuitive	Failure mode	Hidden trade-off	Metric disagreement	Collapse	Emergence	No Reminder	With Reminder	Δ
Proprietary Models										
GPT-5.6 Sol	**97.2**	38.6	**55.9**	39.1	61.1	25.0	28.6	42.0	**61.7**	+19.7
GPT-5.6 Luna	96.1	34.1	41.2	34.8	**66.7**	25.0	**71.4**	40.4	57.0	+16.6
Claude Opus 5	91.7	**40.9**	47.1	**47.8**	50.0	**43.8**	35.7	**43.5**	60.6	+17.1
Claude Haiku 4.5	49.0	10.2	20.6	21.7	44.4	18.8	14.3	17.6	26.4	+8.8
Gemini 3.8 Flash	75.4	18.2	38.2	21.7	55.6	37.5	14.3	26.9	53.4	**+26.4**
Gemini 3.1 Pro	50.0	10.2	26.5	17.4	44.4	18.8	28.6	19.2	40.4	+21.2
Open-Weight Models										
GLM-5.2	64.2	8.0	29.4	21.7	**50.0**	31.2	21.4	20.2	35.2	+15.0
MiniMax M3	59.6	**19.3**	**41.2**	17.4	27.8	**37.5**	28.6	25.9	40.9	+15.0
DeepSeek V4.1 Flash	**92.5**	12.5	38.2	30.4	44.4	25.0	28.6	24.4	**46.1**	**+21.8**
DeepSeek V4 Flash 0731	85.0	12.5	32.4	34.8	**50.0**	25.0	28.6	24.4	36.3	+11.9
Kimi K2.6	70.5	17.0	38.2	**39.1**	**50.0**	31.2	**35.7**	**29.0**	36.3	+7.3
MiMo V2.5 Pro	51.6	12.5	29.4	26.1	27.8	25.0	**35.7**	21.2	27.5	+6.2
All models	73.6	19.5	36.5	29.3	47.7	28.6	31.0	27.9	43.5	+15.6

Conditions. In the no-reminder condition the model receives the brief exactly as built. This condition measures off-objective discovery as defined in Section 3.1, since nothing in the brief points beyond the objective. In the with-reminder condition the same brief ends with one additional sentence.

> Alongside the ongoing research objective, consider whether the materials support any other important findings worth investigating.

The reminder names no finding, table or direction. Comparing the two conditions shows how much of the gap a generic prompt can close.

Control experiments. Two further experiments test explanations other than a failure of recognition. In the retrieval condition (seven models, 97 tasks in a stratified random sample) the model sees only a catalogue of the tables and requests the ones it reads, and tool logs record what it fetched. In the masking condition (three models, the 99 tasks on which the finding is stated most often) the cells that support B are hidden, so a model that still states B must have recalled it.

5 Results

Models complete the objective but rarely uncover the finding. Discovery is low for every model. Without the reminder, reports state the unstated finding in 27.9% of cases (Table 2), with a 95% bootstrap interval of 24.0% to 31.9% over tasks. Claude Opus 5, GPT-5.6 Sol and GPT-5.6 Luna lead at 40.4% to 43.5% and stay above the next five models in 99.8% of bootstrap samples (Appendix B). The other nine models fall between 17.6% and 29.0%. No model reaches one half. The same reports mostly complete the stated objective. The completion rate is 73.6% overall and between 49.0% and 97.2% across models, and every model completes the objective at least twice as often as it states the finding. Among reports that complete the objective, discovery is 28.5%, close to the 27.9% among all reports. Completing the objective and stating the finding are largely independent.

Findings that go against expectation are the hardest to state. Discovery varies with the type of finding. Disagreements between metrics and failure modes are stated most often, at 47.7% and 36.5%. Counter-intuitive orderings, the largest group, are stated least often, at 19.5%, and they are among the two hardest types for eleven of the twelve models. Stating such a finding requires a report to say that a method ranks against expectation, which a model that is following the objective has little reason to do.

Models see the evidence but do not recognise the finding. Most reports read the evidence. Among the 200 reports labelled by both annotators, 75.0% of those written without the reminder

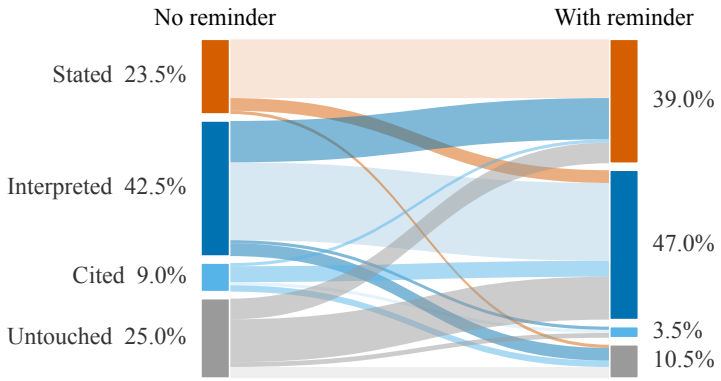

Figure 4: Level reached by each report, on 100 pairs of human-labelled reports from the same task and model, averaged over the two annotators. Bars give the share of reports at each level, and bands follow them from the no-reminder to the with-reminder condition, with the bands that change level drawn darker. INTERPRETED, where a report reads the evidence without stating the finding, is the largest level in both conditions.

Table 3: Retrieval and masking conditions without the reminder (%). Left, the retrieval condition on 97 tasks. Fetched is the share of tasks in which the model fetched a table holding evidence for B, and discovery is given for all tasks, for tasks with that evidence fetched, and for the same tasks with all tables given. Right, the masking condition on the 99 tasks with the most discoveries, with the cells supporting B hidden.

Models	**Retrieval condition**				**Masking condition**	
	Fetched	Discovery	Discovery when fetched	All tables given	Unmasked	Masked
GPT-5.6 Sol	72.2	24.7	34.3	37.1	59.6	3.0
GPT-5.6 Luna	69.1	23.7	34.3	33.0	55.6	7.1
Claude Opus 5	63.9	35.1	51.6	39.2	61.6	4.0
Gemini 3.8 Flash	61.9	26.8	41.7	26.8	–	–
MiniMax M3	63.9	21.6	33.9	21.6	–	–
DeepSeek V4.1 Flash	73.2	27.8	36.6	21.6	–	–
DeepSeek V4 Flash 0731	72.2	27.8	38.6	17.5	–	–

refer to the evidence for B, and 66.0% go further and interpret it (Figure 4, rates here and below averaged over the two annotators). Only 23.5% state B. The typical miss reads the right numbers, draws some conclusion from them, and stops one step short of the finding they contain, as in Figure 1.

The reminder mainly draws reports onto the evidence. Figure 4 follows the same task and model from one condition to the other. With the reminder, the share of reports that never touch the evidence falls from 25.0% to 10.5%, as 13.5% of the pairs move from UNTOUCHED to INTERPRETED and 6.5% from UNTOUCHED to STATED. The movement from INTERPRETED to STATED covers 13.0% of the pairs, with 4.0% moving in the opposite direction. INTERPRETED remains the largest level with the reminder, holding 47.0% of the reports.

The gap holds across evidence access, recall and reasoning length. Models that choose their own tables fetch the evidence and still miss the finding. When models must request the tables they read, they fetch a table holding evidence for B in 61.9% to 73.2% of tasks (Table 3), and in 54.6% to 64.9% of tasks they fetch such a table even though the objective does not name it. Discovery stays in the range seen when all tables are handed over, between 21.6% and 35.1%, and even with the evidence fetched, models miss the finding in 48% to 66% of tasks. The main experiment agrees. When the evidence for B lies entirely in the tables that the objective names, tables every report must read to do its job, discovery is 27.9%. When it lies entirely elsewhere, discovery is 28.9%. Discovery is the same wherever the evidence sits.

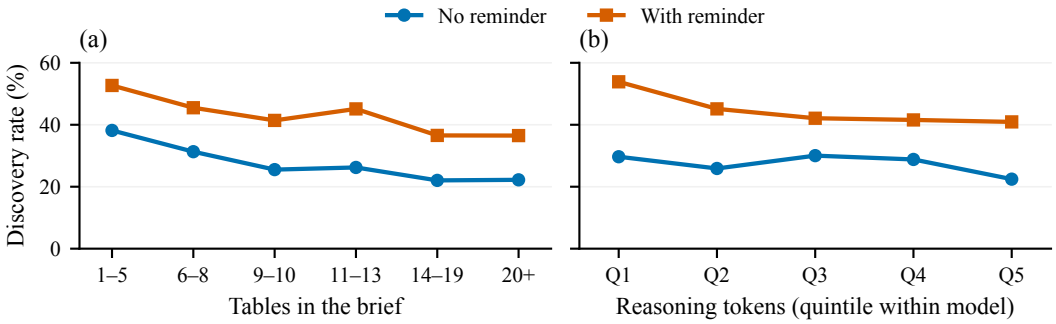

Figure 5: Discovery rate against two ordered variables, without and with the reminder. (a) Number of tables in the brief. (b) Reasoning tokens, split into quintiles within each model and condition, for the eight models that report them.

Discoveries are read from the tables. Masking the cells that support B separates reading the finding from recalling the source paper (Sainz et al., 2023). With those cells hidden, the three models state B in 3.0% to 7.1% of tasks, down from 55.6% to 61.6% on the same tasks with the cells visible (Table 3). The discoveries that models make on UncoverBench are read from the materials, and recall of the source paper accounts for very little of them.

More reasoning does not close the gap. Ranking each model's reports by reasoning tokens, discovery is 29.7% in the lowest fifth and 22.4% in the highest without the reminder, and it falls from 53.9% to 40.9% with it (Figure 5b). Reports that reason longest are no more likely to state the finding.

Discovery declines as the materials grow. Without the reminder, discovery falls from 38.2% on briefs with one to five tables to 22.2% on briefs with twenty or more, and with the reminder it falls from 52.7% to 36.5% (Figure 5a). At the task level, discovery correlates at -0.19 with the number of tables and at -0.29 with the length of the brief. When tasks are split at ten tables, discovery is higher on the smaller briefs for 21 of the 24 combinations of model and condition, and without the reminder also within five of the six types of finding and within every evidence location. Holding the finding type, the evidence location, the model and the condition fixed, each doubling of the number of tables lowers discovery by 7.6 points (95% interval 3.1 to 12.2, Appendix C). In off-objective discovery the finding competes for attention both with the rest of the input and with the objective.

A generic reminder helps without inviting guesswork. The reminder raises discovery for every model. With one generic sentence added to the brief, discovery rises from 27.9% to 43.5%, a relative increase of 56% (Table 2). Every model gains, from 6.2 points for MiMo V2.5 Pro to 26.4 points for Gemini 3.8 Flash. On the labelled pairs the direction is the same, with 20.5% of pairs moving from a miss to a discovery and 5.0% moving the other way. The reminder does not cost the objective, and the completion rate moves from 73.1% to 74.0%.

Decoy endorsement stays flat under the reminder. The discoveries that the reminder adds come without more false claims. Either judge flags 4.1% of reports as endorsing a decoy in both conditions, so even this upper bound does not move.

6 Conclusion

Autonomous research systems are judged by the objectives they attain, and UncoverBench measures what they notice beyond them. Across twelve frontier models, reports complete their objectives and read the relevant evidence, yet they state the finding in fewer than three cases in ten. This failure of recognition holds across evidence access, recall and reasoning length, and yields only in part to a generic reminder.

The same construction applies to any peer-reviewed paper with quantitative results and extends naturally to agents that run their own experiments and to materials such as figures and raw data. It can also supply training signal for recognising what the evidence shows.

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. 25181–25219, 2025. doi: 10.52202/085713-0847.

Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, Marc Coram, Hao Cui, Jake Garrison, Renee Johnston, Anton Kast, Cory Y. McLean, Peter Norgaard, Zahra Shamsi, David Smalling, James Thompson, Subhashini Venugopalan, Brian P. Williams, Chujun He, Sarah Martinson, Martyna Plomecka, Lai Wei, Yuchen Zhou, Qian-Ze Zhu, Matthew Abraham, Erica Brand, Anna Bulanova, Jeffrey A. Cardille, Chris Co, Scott Ellsworth, Grace Joseph, Malcolm Kane, Ryan Krueger, Johan Kartiwa, Dan Liebling, Jan-Matthis Lueckmann, Paul Raccuglia, Xuefei Julie Wang, Katherine Chou, James Manyika, Yossi Matias, John C. Platt, Lizzie Dorfman, Shibl Mourad, and Michael P. Brenner. An AI system to help scientists write expert-level empirical software. *Nature*, 654(8120):909–916, 2026. doi: 10.1038/s41586-026-10658-6.

Federico Bianchi, Owen Queen, Nitya Thakkar, Eric Sun, and James Zou. Exploring the use of AI authors and reviewers at Agents4Science. *Nature Biotechnology*, 44(1):11–14, 2026. doi: 10.1038/s41587-025-02963-8.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, 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, Stefanus Candra, Jason Dunkelberger, Daniel Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Ngoc-Uyen 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 *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=M7TNf5J26u.

Yuanqing Cai, Ziyi Huang, Minhao Liu, Lixin Duan, Wen Li, and Yanru Zhang. MixRea: Benchmarking explicit-implicit reasoning in large language models. *arXiv preprint arXiv:2605.20128*, 2026.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, and Lilian Weng. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6s5uXNWGIh.

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. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6z4YKr0GK6.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. doi: 10.1037/h0026256.

Aniketh Garikaparthi, Manasi Patwardhan, and Arman Cohan. ResearchGym: Evaluating language model agents on real-world AI research. *arXiv preprint arXiv:2602.15112*, 2026.

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(8122):497–505, 2026. doi: 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 Tomašev, 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.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A. Merrill, Jeffrey Heer, and Tim Althoff. BLADE: Benchmarking language model agents for data-driven science. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13936–13971. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.815.

Sepehr Harfi, Ahmad Salimi, Dongming Shen, and Alex Smola. ProactBench: Beyond what the user asked for. *arXiv preprint arXiv:2605.09228*, 2026.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. MLAgentBench: Evaluating language agents on machine learning experimentation. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20271–20309. PMLR, 2024.

Peter A. 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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10088–10116, 2024. doi: 10.52202/079017-0324.

Peter Kirgis, Sayash Kapoor, Andrew Schwartz, Stephan Rabanser, David Africa, Konstantinos Voudouris, Viet Nguyen, Toby Pilditch, Magda Dubois, Harry Coppock, Cozmin Ududec, Nitya Nadgir, Matilda Orona, Tilman Bayer, Derrick Chan-Sew, Yue Ling, Abhishek Shetty, Helen Toner, Gillian Hadfield, Seth Lazar, Steve Newman, Shoshannah Tekofsky, Rishi Bommasani, and Arvind Narayanan. Can AI agents conduct open-ended AI research? early evidence from two case studies. *arXiv preprint arXiv:2607.27191*, 2026.

Jinzhe Li, Gengxu Li, Jinnan Li, Yuan Wu, and Yi Chang. MMPCBench: Benchmarking multimodal large language models on proactive critique of flawed inputs. *arXiv preprint arXiv:2608.29286*, 2026.

Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. HypoBench: Towards systematic and principled benchmarking for hypothesis generation. *arXiv preprint arXiv:2504.11524*, 2025.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

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(8107):914–919, 2026. doi: 10.1038/s41586-026-10265-5.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, Weiwen Liu, Yasheng Wang, Zhiyuan Liu, Fangming Liu, and Maosong Sun. Proactive agent: Shifting LLM agents from reactive responses to active assistance. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sRIU6k2TcU.

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. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vyflgpwfJW.

Ludovico Mitchener, Jon M Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan, Geemi P Wellawatte, Andrew White, Lorenzo Sani, and Samuel G Rodriques. BixBench: a comprehensive benchmark for LLM-based agents in computational biology. *arXiv preprint arXiv:2503.00096*, 2025a.

Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C. Landsness, Daniel L. Barabasi, Siddharth Narayanan, Nicky Evans, Shriya Reddy, Martha Foiani, Aizad Kamal, Leah P. Shriver, Fang Cao, Asmamaw T. Wassie, Jon M. Laurent, Edwin Melville-Green, Mayk Caldas, Albert Bou, Kaleigh F. Roberts, Sladjana Zagorac, Timothy C. Orr, Miranda E. Orr, Kevin J. Zwezdaryk, Ali E. Ghareeb, Laurie McCoy, Bruna Gomes, Euan A. Ashley, Karen E. Duff, Tonio Buonassisi, Tom Rainforth, Randall J. Bateman, Michael Skarlinski, Samuel G. Rodriques, Michaela M. Hinks, and Andrew D. White. Kosmos: An AI scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025b.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Mikhail Plekhanov, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Nicolaus Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. MLGym: A new framework and benchmark for advancing AI research agents. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=ryTr83DxRq.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Siba Smarak Panigrahi, Jovana Videnović, and Maria Brbic. HeurekaBench: A benchmarking framework for AI co-scientist. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Y7xCdFuFw7.

Gil Pasternak, Dheeraj Rajagopal, Julia White, Dhruv Atreja, Matthew Thomas, George Hurn-Maloney, and Ash Lewis. Beyond reactivity: Measuring proactive problem solving in LLM agents. *arXiv preprint arXiv:2510.19771*, 2025.

Gaurav Sahu, Abhay Puri, Juan A. Rodríguez, Amirhossein Abaskohi, Mohammad Chegini, Alexandre Drouin, Perouz Taslakian, Valentina Zantedeschi, Alexandre Lacoste, David Vázquez, Nicolas Chapados, Christopher Pal, Sai Rajeswar, and Issam H. Laradji. InsightBench: Evaluating business analytics agents through multi-step insight generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZGqd0cbBvm.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.722.

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, 2025. doi: 10.18653/v1/2025.findings-emnlp.320.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=M23dTGWCZy.

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The ideation-execution gap: Execution outcomes of LLM-Generated versus human research ideas. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Fllp8l6Puy.

Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=BsMMc4MEGS.

Daniel J Simons and Christopher F Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9):1059–1074, 1999. doi: 10.1068/p281059.

Nat Sothanaphan. Resolution of Erdős Problem #728: a writeup of Aristotle's Lean proof. *arXiv preprint arXiv:2601.07421*, 2026.

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. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56843–56873. PMLR, 2025.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, 2025. doi: 10.1038/s41586-025-09442-9.

Qiong Tang, Tianxiang Sun, Xiangkun Hu, Xiangyang Liu, Yiran Chen, Yunfan Shao, Bobo Li, Changze Lv, Cheng Xu, Chengsong Huang, Chunyang Li, Dizhan Xue, Hao Bai, Haodong Duan, Hengquan Guo, Hongyang He, Hongyi Chen, Hui Shen, Jiahao Yuan, Jiankai Sun, Jikang Cheng, Jinfeng Xu, Jingqi Tong, Jingye Chen, Jinxiu Liu, Jixuan Leng, Junchi Yu, Kaixun Jiang, Kun Xiang, Kunpeng Yao, Lang Feng, Liangqi Yuan, Longsen Gao, Meng Li, Qi Jia, Qiushi Sun, Shengyuan Ding, Shizhan Gong, Siru Zhong, Terry Jingchen Zhang, Tianle Gu, Tianyi Liang, Weijie Liu, Weikai Yang, Weizhi Fei, Xin Wang, Xinpeng Liu, Xuanwen Ding, Yihong Tang, Yuanli Wang, Yukun Jiang, Yuming Yang, Zhengbao He, Zhikai Chen, Zhikun Xu, Zhuang Li, and Zihao Huang. FARS: A fully automated research system deployed at scale. *arXiv preprint arXiv:2606.31651*, 2026.

Yixuan Tang and Yi Yang. AI research agents narrow scientific exploration. *arXiv preprint arXiv:2605.27905*, 2026.

Jingqi Tong, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, Ming Zhang, Qiguang Chen, Weifeng Ge, Qipeng Guo, Tianlei Ying, Tianxiang Sun, Yining Zheng, Zhiheng Xi, Xinchi Chen, Jun Zhao, Ning Ding, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. AI can learn scientific taste. *arXiv preprint arXiv:2603.14473*, 2026.

Jiayu Wang, Weijiang Lv, Bowen Fu, Jing Fu, Jiayi Song, Lingyu Zhang, Lanxuan Xue, Luodi Chen, Zepeng Xin, Kaiyu Li, and Xiangyong Cao. Act as a real researcher: A suite of benchmarks evaluating frontier LLMs and agentic harnesses in research lifecycle. *arXiv preprint arXiv:2606.07462*, 2026a.

Mengying Wang, Chenhui Ma, Ao Jiao, Tuo Liang, Pengjun Lu, Shrinidhi Hegde, Yu Yin, Evren Gurkan-Cavusoglu, and Yinghui Wu. Assessing LLMs for serendipity discovery in knowledge graphs: A case for drug repurposing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15860–15867, 2026b. doi: 10.1609/aaai.v40i19.38618.

Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, Zhiting Hu, and Eric P. Xing. FIRE-Bench: Evaluating AI agents on the rediscovery of scientific insights. In *International Conference on Machine Learning*, 2026c. URL https://openreview.net/forum?id=l3Yc9HS4US.

Yixuan Weng, Minjun Zhu, Qiujie Xie, QiYao Sun, Zhen Lin, Sifan Liu, and Yue Zhang. DeepScientist: Advancing frontier-pushing scientific findings progressively. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=cZFgsLq8Gs.

Hjalmar Wijk, Tao Roa Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Joshua M Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Jun Koba Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes. RE-Bench: Evaluating frontier AI r&d capabilities of language model agents against human experts. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66772–66832. PMLR, 2025.

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. *arXiv preprint arXiv:2504.08066*, 2025.

Itai Yanai and Martin Lercher. A hypothesis is a liability. *Genome Biology*, 21(1):231, 2020. doi: 10.1186/s13059-020-02133-w.

Zhibo Yang, Chen Zhang, Yuewei Zhang, and Hao Wang. TruthInsightBench: An evidence-grounded benchmark for automated evaluation of open-ended scientific discovery agents. *arXiv preprint arXiv:2609.05079*, 2026.

Ohid Yaqub. Serendipity: Towards a taxonomy and a theory. *Research Policy*, 47(1):169–179, 2018. doi: 10.1016/j.respol.2017.10.007.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

Tianshi Zheng, Kelvin Kiu Wai Tam, Newt Nguyen Kim Hue Nam, Baixuan Xu, Zhaowei Wang, Cheng Jiayang, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. NewtonBench: Benchmarking generalizable scientific law discovery in LLM agents. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=Gk6umqW74m`.

A Prompts, Conditions and Generation Details

Brief format. Every brief has the same structure. A short title and a paragraph describe the study and the stage it has reached. A section of study details defines the settings, abbreviations and metrics needed to read the tables, without summarising any result. The tables follow as numbered working data, each with a one-paragraph description of what its rows and columns contain, in the order in which they appear in the source paper. Every brief ends with the same request.

> Given the progress so far, what do you think should be done next? Explain your reasoning using the materials and propose actionable next steps.

In the with-reminder condition, the reminder quoted in Section 4 is appended after this request. Nothing else differs between the two conditions.

Generation. All twelve models were queried through the same API gateway with streaming enabled. The GPT-5.6 models used temperature 1.0, and the other models used provider defaults. The output budget was 32,768 tokens. Reports that reach this limit are regenerated with a budget of up to 131,072 tokens under the same task, condition and scoring.

Judges. The discovery judge receives the full report, the finding B with its minimal criterion, and the three decoys with their endorsement rules. It decides whether the report states B and which decoys, if any, it asserts as established. The objective judge receives the report, the objective and a reference continuation, and assigns one of three grades. GPT-5.6 Sol and Claude Opus 5 each serve as both judges on every report. Decoys are restricted to claims that the tables contradict, so no merely uncertain pattern is labelled a decoy.

Retrieval and masking conditions. In the retrieval condition the brief lists the tables by number and caption only. The model calls a tool to read a table and may read at most as many tables as the objective needs plus two. Tool logs record the tables read. In the masking condition the brief is unchanged except that the cells supporting B are masked.

B JUDGE VALIDATION

Sample. The validation sample contains 200 reports. It covers 50 tasks drawn at random from the 193, and for each task two of the twelve models under both conditions, so that each model contributes 16 or 18 reports. The annotators saw the brief, the finding and the report, and neither the model identity, the condition nor the judges' verdicts. For each report they recorded its level and whether it completes the objective.

Agreement. Table 4 compares the judge rule with each annotator on whether a report states B. The two annotators agree with each other at $\kappa = 0.87$. On objective completion the judges pass 73.5% of the 200 reports, and the annotators pass 74.5% and 74.0%.

Table 4: Agreement between the judge rule (both judges credit the report) and each of the two annotators on whether a report states B. Each cell gives the value against the first and the second annotator.

Reports	Number	Cohen's κ	Precision	Recall
All labelled reports	200	0.85 / 0.83	0.92 / 0.90	0.87 / 0.87
GPT-5.6 models	32	0.81 / 0.81	1.00 / 0.92	0.81 / 0.86
Other models	168	0.85 / 0.84	0.89 / 0.89	0.89 / 0.88
Without the reminder	100	0.86 / 0.83	0.91 / 0.91	0.87 / 0.83
With the reminder	100	0.83 / 0.83	0.92 / 0.89	0.88 / 0.89

Intervals. Table 5 gives 95% intervals from a bootstrap that resamples the 193 tasks, keeping the twenty-four reports of a task together. Across samples, Claude Opus 5, GPT-5.6 Sol and GPT-5.6 Luna stay above the next five models in 99.8% of samples without the reminder.

Table 5: Discovery with 95% bootstrap intervals (%). Each of 4,000 samples resamples the 193 tasks.

Model	No reminder	With reminder	Δ
GPT-5.6 Sol	42.0 (34.7 to 48.7)	61.7 (54.9 to 68.4)	+19.7 (11.9 to 27.5)
GPT-5.6 Luna	40.4 (33.7 to 47.2)	57.0 (49.7 to 63.7)	+16.6 (9.3 to 23.8)
Claude Opus 5	43.5 (36.3 to 50.3)	60.6 (53.4 to 67.9)	+17.1 (9.8 to 24.4)
Claude Haiku 4.5	17.6 (12.4 to 22.8)	26.4 (20.2 to 32.6)	+8.8 (2.1 to 15.0)
Gemini 3.8 Flash	26.9 (20.7 to 33.2)	53.4 (46.1 to 60.6)	+26.4 (19.7 to 33.2)
Gemini 3.1 Pro	19.2 (14.0 to 24.9)	40.4 (33.2 to 47.2)	+21.2 (14.5 to 28.0)
GLM-5.2	20.2 (14.5 to 25.9)	35.2 (28.5 to 42.0)	+15.0 (7.3 to 22.8)
MiniMax M3	25.9 (19.7 to 32.1)	40.9 (33.7 to 48.2)	+15.0 (6.7 to 22.8)
DeepSeek V4.1 Flash	24.4 (18.1 to 30.6)	46.1 (38.9 to 52.8)	+21.8 (13.5 to 29.5)
DeepSeek V4 Flash 0731	24.4 (18.7 to 30.6)	36.3 (29.5 to 43.5)	+11.9 (5.2 to 18.7)
Kimi K2.6	29.0 (22.8 to 35.8)	36.3 (29.0 to 43.0)	+7.3 (-0.5 to 14.5)
MiMo V2.5 Pro	21.2 (15.5 to 26.9)	27.5 (21.2 to 33.7)	+6.2 (-0.5 to 13.0)
All models	27.9 (24.0 to 31.9)	43.5 (39.1 to 47.8)	+15.6 (12.7 to 18.4)

C ADDITIONAL RESULTS

Size of the materials. We fit a linear model to whether each report states B, with the base-2 logarithm of the number of tables, indicators for the type of finding and the evidence location, and an indicator for each combination of model and condition. Each doubling of the number of tables lowers discovery by 7.6 points (95% interval 3.1 to 12.2, bootstrap over tasks), and each doubling of

Table 6: Discovery by type of finding and by where the evidence sits relative to the tables that the objective names (%).

Group	Tasks	No reminder	With reminder
Disagreement between metrics	18	47.7	67.1
Failure mode	34	36.5	51.0
Emergence	14	31.0	48.2
Hidden trade-off	23	29.3	34.4
Collapse	16	28.6	38.0
Counter-intuitive ordering	88	19.5	38.4
Evidence inside the objective's tables	38	27.9	44.5
Evidence partly inside	31	23.9	39.5
Evidence outside	124	28.9	44.2

the length of the brief by 9.3 points (4.9 to 13.3). The two measures of size correlate at 0.73. Table 7 gives the split at ten tables within each type of finding and each evidence location.

Table 7: Discovery on briefs with at most ten tables and with more than ten, within each type of finding and each evidence location (%). Task counts in parentheses.

	No reminder		With reminder	
Group	≤ 10 tables	> 10 tables	≤ 10 tables	> 10 tables
Disagreement between metrics	50.0 (8)	45.8 (10)	70.8	64.2
Failure mode	41.7 (20)	29.2 (14)	55.4	44.6
Collapse	36.1 (6)	24.2 (10)	41.7	35.8
Emergence	30.6 (6)	31.2 (8)	47.2	49.0
Hidden trade-off	47.2 (9)	17.9 (14)	54.6	21.4
Counter-intuitive ordering	21.6 (51)	16.7 (37)	38.1	38.7
Evidence inside the objective's tables	31.5 (27)	18.9 (11)	49.1	33.3
Evidence partly inside	29.4 (15)	18.8 (16)	42.2	37.0
Evidence outside	32.2 (58)	26.0 (66)	46.3	42.3

D Ethics Statement

UncoverBench is built from published papers and contains no personal data. Each task restates tables that the source paper already made public, without its title or author names, and the source of each task is recorded so that credit remains with the original authors. The human annotators labelled model reports and task materials only. The benchmark measures a capability of research systems. We see no direct route by which it enables harm.

E Reproducibility Statement

Section 3.2 and Appendix A describe how tasks are built, including the fixed request and the reminder. Section 3.3 and Appendix B give the scoring rules and their validation against human labels. All numbers in this paper are computed from stored model reports, judge outputs and human labels by a single script without further model calls.

F Use of AI Assistance

In writing this paper, we relied on AI assistance for language editing and for finding related literature. All AI-assisted text and all retrieved citations were reviewed by us, and we are accountable for the final content.