MECHHYPOBENCH: EVALUATING DATA-DRIVEN MECHANISTIC HYPOTHESIS GENERATION IN AI AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-driven mechanistic hypotheses are essential to scientific discovery because they explain how underlying processes produce observed phenomena. AI agents and AI scientists increasingly support scientific data analysis. However, their ability to turn empirical findings into mechanistic hypotheses remains insufficiently examined. To address this gap, we introduce MechHypoBench, the first benchmark for evaluating whether AI agents and AI scientists can generate such hypotheses from empirical data. It combines paper-derived mechanisms from 14 scientific fields with real-world datasets containing 17.98 million records. The construction retains the observational complexity of empirical data while providing a specified underlying mechanism. Agents analyze the observations and propose open-form hypotheses. We develop an evaluation framework that assesses open-form mechanistic hypotheses through their consequence under withheld conditions. Experiments with general agents and AI scientists reveal a substantial gap between generated hypotheses and the underlying mechanisms. **Our datasets and codes are available** at `https://anonymous.4open.science/r/ICLR-2027-benchmark-E1FB`.

1 INTRODUCTION

Recent advances in AI agents and AI scientists have largely accelerated scientific discoveries across domains (Wang et al., 2023a; Boiko et al., 2023; Yamada et al., 2025; Gottweis et al., 2026), including physics, biology, chemistry, etc. One of the central tasks is to use observed data to investigate the underlying mechanism of how scientific phenomena arise. For example, when a catalytic reaction slows over time, researchers may hypothesize that molecules progressively block the catalyst's active sites, leaving fewer sites available for reaction. A mechanistic hypothesis is an open-form account of how processes and their interactions produce an observed phenomenon, including the conditions under which they operate. In our review of papers from natural-science journals, 70.2% treated a mechanism as an explanatory output (Figure 1). Assessing AI agents' ability to generate mechanistic hypotheses from data is therefore important for determining whether they can contribute to scientific discovery.

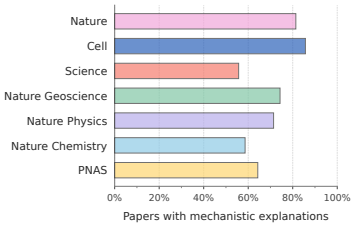


Figure 1: We randomly select papers published in 2021-2025 from high-impact journals (Appendix C.1). Overall, 70.2% of the papers treated mechanisms as explanatory outputs, motivating our research on mechanism-driven hypothesis generation.

Existing benchmarks have expanded the evaluation of LLMs in scientific discovery beyond question answering (Jin et al., 2019; Wang et al., 2023b). Several benchmarks study literature-grounded hypothesis and idea generation using scientific papers and background knowledge (Guo et al., 2025; Liu et al., 2026; Lew et al., 2026). Others evaluate hypothesis discovery from specified entities or scientific data, including the recovery of entity relations, equations, data relationships, explanatory patterns, and admissible hypothesis sets (Xiong et al., 2025; Majumder et al., 2025; Shojaee et al., 2025; Liu et al., 2025; Chen et al., 2026). A separate group evaluates scientific research and discovery workflows, ranging from executable data-analysis to broader task- and project-level evaluations (Chen et al., 2025; Bragg et al., 2026; Song et al., 2025).

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.

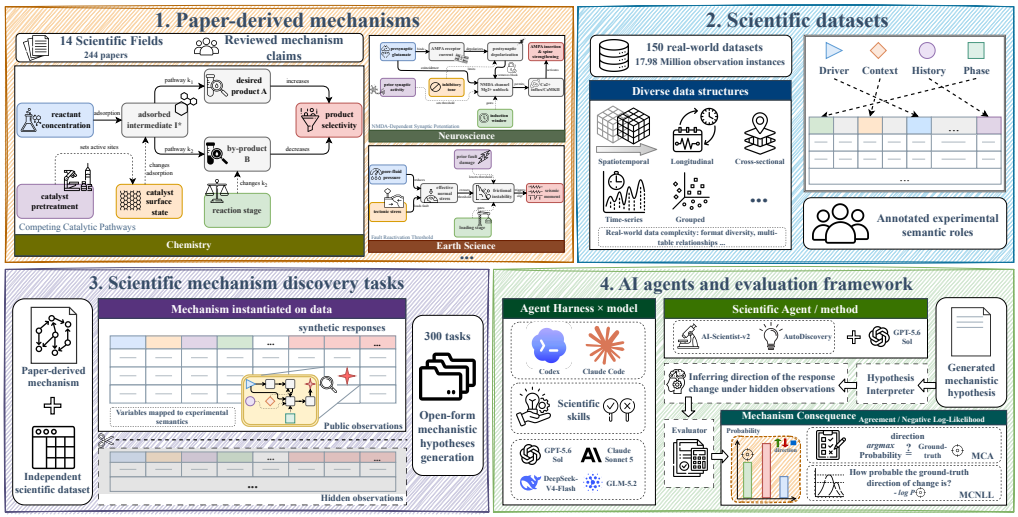


Figure 2: Overview of our MechHypoBench.

However, existing benchmarks pose unique challenges for evaluating data-driven mechanistic hypotheses. **(1) Finding a pattern does not explain how a phenomenon arises.** Most data-driven benchmarks ask models to recover relationships or equations (Majumder et al., 2025; Shojaee et al., 2025). Such tasks show whether a model can detect what happens in the data, but not whether it can explain how underlying processes produce the phenomenon under specific conditions. The *Mechanistic* column of Table 1 shows that direct evaluation of this ability remains limited, even among benchmarks that accept explanatory hypotheses. **(2) Known mechanisms rarely come with realistic observations.** Real datasets preserve the complexity of scientific observations, but they rarely identify a unique and fully specified mechanism (Liu et al., 2025). Treating a source-paper explanation as ground truth may privilege one interpretation and reward familiarity with the paper rather than reasoning from the data. Synthetic data provide exact references but simplify the observations (Chen et al., 2026). Accordingly, the *Exp. Data* and *Source* columns of Table 1 show that no prior benchmark fully combines realistic empirical data with a controllable underlying mechanism. **(3) Open-form hypotheses cannot be scored by wording alone.** Equivalent mechanisms may use different wording, while a hypothesis may have clear implications only under some conditions. Text matching or a single classification decision cannot capture these differences. Evaluation should instead test the outcomes implied under unseen conditions to assess how well a hypothesis captures the behavior of the underlying mechanism. The pipeline should also support agents that iteratively inspect data and execute analyses. The *Open-form* and *Agent* columns of Table 1 show that existing benchmarks rarely provide both.

In response, we introduce MechHypoBench, a benchmark for data-driven mechanistic hypothesis generation with AI agents (Figure 2). We instantiate mechanisms derived from papers spanning 14 scientific fields (Priem et al., 2022) on independently collected real-world scientific datasets. The datasets vary in structure and collectively contain 17.98 million observation records. This construction combines a scientifically grounded generating process with empirical observational complexity by aligning mechanisms and datasets through their experimental roles. Agents analyze the resulting observations, then propose open-form mechanistic hypotheses to explain how the observed phenomena arise. We assess these hypotheses through their implied consequences under withheld changes in conditions, using ground truth outcomes determined by the same generating mechanism. Our experiments reveal substantial room for improvement in agents' ability to formulate mechanistic hypotheses from scientific data. **We summarize our contributions as follows:**

- We propose MechHypoBench, the first benchmark for data-driven mechanistic hypothesis generation. It combines paper-derived mechanisms from 14 scientific fields with independently collected empirical datasets.
- We develop an evaluation framework for open-form mechanistic hypotheses. It assesses how well a hypothesis captures the behavior of the underlying mechanism.

Table 1: Comparison of MechHypoBench with existing scientific discovery benchmarks. ∼ denotes partial coverage. # Sci. Subj. denotes the number of real scientific fields covered by each work.

Type	Work	# Sci. Subj.	Task Source		Hypothesis		Agent Pipeline		Task Scale
			Exp. Data	Source	Mechanistic	Open-form	Skill	Agent	
Literature -Driven	IdeaBench (Guo et al., 2025)	1	✗	papers	✗	✓	✗	✗	unknown
	ResearchBench (Liu et al., 2026)	12	✗	papers	✗	✓	✗	✗	unknown
	ProjectionBench (Lew et al., 2026)	1	✗	papers	✗	✓	✗	✗	unknown
	TruthHypo (Xiong et al., 2025)	1	✗	Bio. KG	✗	✗	✗	✗	2k tasks
Research Workflow	ScienceAgentBench (Chen et al., 2025)	4	✓	44 (paper, dataset, code)	✗	✗	✗	✓	102 tasks
	AstaBench (Bragg et al., 2026)	unknown	∼	11 benchmarks	✗	∼	✗	✓	2.4k tasks
	SDE (Song et al., 2025)	4	∼	43 scenarios	∼	✗	✗	✗	1k QA + 8 proj.
Data-Driven	DiscoveryBench (Majumder et al., 2025)	6	∼	27 (paper, dataset, code)	✗	✓	✗	✓	1k tasks
	LLM-SRBench (Shojaee et al., 2025)	4	✗	equations	✓	✗	✗	✗	239 tasks
	HypoBench (Liu et al., 2025)	7	∼	datasets	∼	✓	✗	✗	unknown
	HypoSpace (Chen et al., 2026)	unknown	✗	Synth.	✗	✗	✗	✗	unknown
	Ours	**14**	✓	**244 papers + 150 datasets**	✓	✓	✓	✓	**300 tasks**

- Through experiments with AI scientists, we identify limitations in data-driven mechanistic hypothesis generation. A substantial gap remains between the generated hypotheses and the underlying mechanisms.

2 Related Work

We position MechHypoBench among existing benchmarks for scientific reasoning and discovery in Table 1. These benchmarks extend beyond question answering (Jin et al., 2019; Wang et al., 2023b). Literature- and knowledge-driven benchmarks evaluate idea or hypothesis generation from published research and background knowledge (Guo et al., 2025; Liu et al., 2026; Lew et al., 2026; Xiong et al., 2025). But discovery also requires direct analysis of observational data to establish empirical support for these proposals. Existing data-driven benchmarks evaluate discovery and hypothesis generation through the recovery of relationships, equations, ground-truth hypotheses, and members of enumerable hypothesis spaces (Majumder et al., 2025; Shojaee et al., 2025; Liu et al., 2025; Chen et al., 2026). However, their empirical settings generally lack fully specified underlying mechanisms. Controlled synthetic settings, in turn, simplify the complexity of real scientific observations. The growing use of AI agents also motivates evaluating how they conduct analyses and interact with tools in discovery. Research-workflow benchmarks focus on executable analyses and end-to-end scientific tasks (Chen et al., 2025; Bragg et al., 2026; Song et al., 2025). Yet successful task execution does not necessarily imply successful mechanistic reasoning. MechHypoBench fills this gap by directly evaluating AI agents' ability to generate open-form mechanistic hypotheses from data while retaining the observational complexity of empirical scientific datasets.

3 MechHypoBench

In this section, we first define the task, then describe scientific source collection and benchmark construction, and finally present the evaluation framework.

3.1 Concept Formalization

Mechanisms, conditions, and observations. Scientific inquiry draws on observed data collected under particular conditions to investigate how phenomena arise (Woodward, 2011). An **underlying mechanism** $\mathcal{M}$ generates the phenomenon, but the available observations may reveal only part of how it operates. Its observable response may depend not only on the factor being examined, but also

on the surrounding context, prior exposure, and the stage of the process (Machamer et al., 2000). To represent these dependencies, we distinguish four **experimental semantic roles**:

- **Driver (**d**):** the factor whose effect is examined.
- **Context (**x**):** a contextual factor that may modify the driver's effect.
- **History** (h)**:** prior exposure or state.
- **Phase (**p**):** the stage of the process.

A condition $c = (c_{\mathrm{d}}, c_{\mathrm{x}}, c_{\mathrm{h}}, c_{\mathrm{p}}) \in \mathcal{C}$ specifies the joint values or states associated with these roles, where $\mathcal{C}$ denotes the relevant condition space. Let $R_{\mathcal{M}}(c)$ denote the response, the observable outcome generated by mechanism $\mathcal{M}$ under condition c. For instance, in the chemistry example of Figure 2, c_{d} specifies the *reactant concentration*, c_{x} the *catalyst surface state*, c_{h} the *catalyst pretreatment*, and c_{p} the *reaction stage*. Together, these specify a condition c, with *product selectivity* as the response $R_{\mathcal{M}}(c)$. Changing the *reactant concentration* c_{d} while keeping the other three components fixed yields a different condition. Let $\mathcal{C}^{\mathrm{obs}} \subseteq \mathcal{C}$ denote the conditions under which observations are available. These observations can be represented as condition-response pairs: $\mathcal{O}_{\mathcal{M}}(\mathcal{C}^{\mathrm{obs}}) = \{(c, R_{\mathcal{M}}(c)) \mid c \in \mathcal{C}^{\mathrm{obs}}\}$. In the chemistry example, $\mathcal{C}^{\mathrm{obs}}$ contains the combinations of *concentration, catalyst state, pretreatment*, and *reaction stage* under which measurements were collected; $\mathcal{O}_{\mathcal{M}}(\mathcal{C}^{\mathrm{obs}})$ pairs those combinations with their corresponding *product selectivity*.

Mechanistic hypothesis. Since the available observations do not fully reveal how the underlying mechanism operates, researchers formulate mechanistic hypotheses to explain how the observed phenomenon arises (Darden, 2002; Craver & Darden, 2013). A mechanistic hypothesis h is an open-form account of how relevant processes and their interactions produce the observations $\mathcal{O}_{\mathcal{M}}(\mathcal{C}^{\mathrm{obs}})$, including the conditions under which they operate. The available observations may support multiple mechanistic hypotheses to differing degrees.

3.2 Scientific Source Collection

We collect scientific papers to derive mechanisms and independently gather real-world empirical datasets on which these mechanisms are instantiated.

Paper collection. We collected candidate papers through Europe PMC and OpenAlex(Priem et al., 2022) across a range of scientific journals, including PLOS journals and eLife. **(1) Candidate filtering.** We retained empirical studies that proposed, tested, or revised mechanisms, excluding reviews, protocols, retracted papers, and studies focused solely on prediction or optimization. **(2) Mechanistic claim review.** Human review assessed whether the reported claims described how a phenomenon arises and provided sufficient information to identify the relevant experimental semantic roles and conditions. **(3) Claim extraction and verification.** We used LLMs to assist in extracting mechanistic claims and supporting passages, then manually checked the extracted content against the source papers, preserving the claims and their conditions of applicability. The resulting collection contains 306 papers spanning 14 scientific fields.

Dataset collection. We independently collected candidate datasets from scientific repositories, including PANGAEA (Felden et al., 2023), ScienceDB (Long et al., 2026), Zenodo (Sicilia et al., 2017), Dryad (He & Nahar, 2016), and the Open Science Framework (Foster & Deardorff, 2017). **(1) Candidate filtering.** We selected datasets with traceable provenance, accessible data files, suitable reuse licenses, and documentation describing the recorded variables. **(2) Scientific and structural review.** Human review assessed whether each dataset contained sufficient variation, ordered observations, auxiliary measurements, and grouping structure to support the experimental semantic roles and the intended task construction. **(3) Dataset preparation and role annotation.** Dataset variables were manually annotated with their corresponding experimental semantic roles. Dataset-specific preprocessing adapters accommodated differences in file formats and table layouts. The resulting collection contains 200 datasets with manual role annotations and diverse structures, including cross-sectional, longitudinal, time-series, grouped, and spatiotemporal formats, among others.

3.3 Scientific Skills Collection

To examine whether scientific skills affects mechanistic hypothesis generation, we collected over 500,000 science-related skills from GitHub and filtered them for relevance to scientific data analysis,

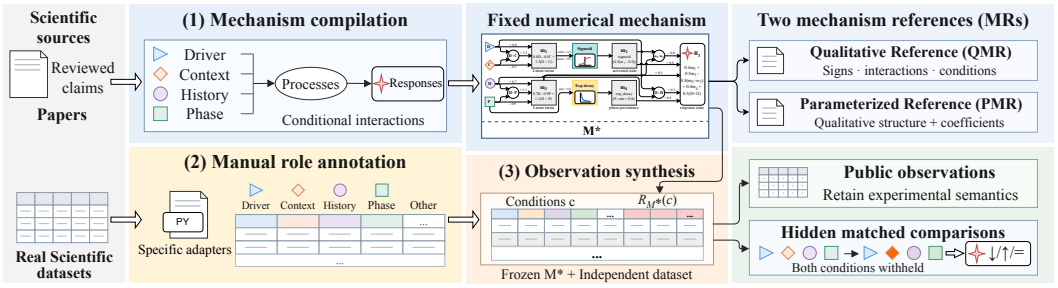


Figure 3: Task construction pipeline. Paper-derived claims are compiled into a fixed mechanism $\mathcal{M}^*$ and aligned with independent scientific datasets through experimental roles. The mechanism generates responses on the aligned conditions and yields two open-form references. Agents receive public observation and are evaluated on hidden matched comparisons.

discovery, and hypothesis generation. **(1) Initial filtering.** An LLM used for screening skill descriptions to shortlist over 20,000 candidates. Filtering excluded academic writing and formatting tools, literature-retrieval and web-dependent tools, and generic agent frameworks or LLM API wrappers. **(2) Relevance review.** Further LLM-based review examined both descriptions and skill contents to identify candidates relevant to scientific data analysis, discovery, or hypothesis generation. **(3) Pool consolidation.** After removing duplicates and entries with insufficient content, the final pool contained 4,043 skills.

3.4 Task Construction

Figure 3 illustrates how we simulate the scientific inference setting using a specified **paper-derived mechanism** $\mathcal{M}^*$, which serves as the underlying mechanism within each task. It is known to the benchmark designer but hidden from the agent. We use LLM-assisted conversion to organize reviewed paper claims into a directed graph of processes, interactions, and responses associated with the experimental semantic roles. Numerical functions and parameters specify how these components operate. The resulting mechanism is checked against the reviewed claims and fixed before dataset pairing, providing a controlled realization of the reported processes rather than a numerical reconstruction of the original experiment. From the same mechanism graph, we derive two open-form textual **mechanism references** (MRs): the **qualitative mechanism reference** (QMR) describes its dependencies, interactions, and conditions, while the **parameterized mechanism reference** (PMR) additionally provides the corresponding equations and numerical parameters.

We align an independently collected dataset with $\mathcal{M}^*$ through its manually annotated experimental semantic roles. Dataset variables specify conditions c, under which the mechanism generates responses $R_{\mathcal{M}^*}(c)$. This preserves empirical sampling patterns, grouping, and auxiliary measurements, with added noise representing measurement variability. Dataset labels and task descriptions retain the experimental roles and their interpretation while replacing source-specific terminology to avoid misleading domain associations and cues to memorized scientific accounts. The generated observations are divided into **public observations** $\mathcal{O}_{\mathcal{M}^*}(\mathcal{C}^{\text{obs}})$, available to the agent, and **hidden observations**, withheld for assessment. Each task includes a set $\mathcal{Q}$ of **matched comparisons** $q = (c^{(1)}, c^{(0)}, R)$, each comparing response R under two conditions that differ in exactly one component. These comparisons have stable ground-truth direction of response changes determined by $\mathcal{M}^*$ and sufficient observational support. Observations under both compared conditions belong to the hidden portion, although individual component values may still occur in the public observations. Task difficulty is determined by a marginal baseline fitted to the public observations: easy tasks yield one to three errors across ten directional comparisons, whereas hard tasks yield six to ten, distinguishing how poorly marginal relationships capture the conditional response changes.

3.5 Human Evaluation

To assess the quality of the constructed tasks, we conducted a human evaluation across both difficulty levels. Ten reviewers with scientific backgrounds assessed each task on four dimensions using a 0-5 scale: (1) **Scientific Coherence** (SC) assesses whether the mechanism, experimental roles, and

comparisons form a scientifically interpretable scenario; (2) **Mechanism-task Consistency** (MC) measures agreement among the mechanism, role mapping, data-generation process, and comparisons; (3) **Probe Clarity** (PC) assesses whether the condition change, response variable, and requested judgment are unambiguous; and (4) **Reference-Answer Validity** (RV) assesses how clearly the ground-truth direction of response changes are supported by the specified mechanism.

Human review prompted the reconstruction of 17 paper-derived mechanisms with insufficient scientific coherence. Table 2 reports scores after these revisions. High ratings for probe clarity (4.90) and reference-answer validity (4.96-4.97) support the clarity of the comparisons and the validity of their ground-truth response changes. Scientific coherence (4.60-4.62) and mechanism-task consistency (4.54-4.69) indicate that the revised tasks present interpretable scientific scenarios with internally consistent construction.

Table 2: Human evaluation scores after mechanism reconstruction (0-5).

Task	SC	MC	PC	RV
Easy	4.62	4.69	4.90	4.97
Hard	4.60	4.54	4.90	4.96

3.6 Evaluation Framework

Evaluation Framework. Mechanistic hypotheses support reasoning about interventions and outcomes under conditions not yet observed (Woodward & Woodward, 2005; Craver, 2007). Motivated by this perspective, we assess hypotheses through the response changes they imply for the hidden matched comparisons. For each $q = (c^{(1)}, c^{(0)}, R)$, let $\Delta_{\mathcal{M}^*}(q)$ denote the ground-truth direction of change in response R from $c^{(0)}$ to $c^{(1)}$ determined by the paper-derived mechanism, and let $\Delta_h(q)$ denote the corresponding direction implied by hypothesis h. Agreement between these directions provides a criterion for assessing the hypothesis. Open-form mechanistic hypotheses do not necessarily specify an executable response function, so their implications must be interpreted before scoring. Asking the generating agent to answer the comparisons could draw on observations or intermediate analyses not expressed in its hypothesis. We therefore use a separate, fixed LLM-based **hypotheses interpreter** that receives only h and the comparison descriptions, without access to the public observations or ground-truth directions. For each comparison q, it returns $p_h(t \mid q)$ over $\mathcal{T} = \{\text{decrease}, \text{no_change}, \text{increase}\}$, representing its interpretation of $\Delta_h(q)$. We request probabilities rather than a single direction because a hypothesis may leave some consequences ambiguous or incompletely specified. A probability distribution allows the interpreter to express this uncertainty instead of forcing an equally definite answer for every hypothesis. A deterministic evaluator then computes scores against $\Delta_{\mathcal{M}^*}(q)$. The same interpreter processes both MRs, providing a comparison with specified underlying mechanism content in each task. Empty and shuffled controls provide baselines for absent and mismatched content, respectively.

Metrics. We assess each mechanistic hypothesis using two complementary metrics:

$$\begin{aligned}\mathrm{MCA}(h) &= \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathbf{1}\left[\arg\max_{t\in\mathcal{T}} p_h(t \mid q) = \Delta_{\mathcal{M}^*}(q)\right], \\ \mathrm{MCNLL}(h) &= -\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \log p_h(\Delta_{\mathcal{M}^*}(q) \mid q)\,,\end{aligned}$$

where Mechanism Consequence Agreement (MCA) measures how consistently the hypothesis-implied direction of response change agree with those determined by the generating mechanism. Mechanism Consequence Negative Log-Likelihood (MCNLL) accounts for the probability assigned to the ground-truth direction, penalizing confident errors more strongly. Together, higher MCA and lower MCNLL provide evidence that a mechanistic hypothesis captures the behavior of the underlying mechanism under the tested conditions.

4 Experiments

4.1 Experimental Setup

Tasks. Our evaluation includes 300 tasks, drawing on 244 papers and 150 datasets from source collections of 306 papers and 200 datasets (Section 3.2). It comprises 120 easy and 180 hard tasks.

Table 3: Performance of mechanistic hypothesis generation evaluation. QMR and PMR denote qualitative and parameterized mechanism references. Empty and Shuffled denote absent and mismatched hypothesis controls.

Type	Models	Without Skills				With Skills			
		MCA (%)↑		MCNLL ↓		MCA (%)↑		MCNLL ↓	
		Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard
Ref. & Controls	*PMR*	**96.8**	**97.5**	**0.165**	**0.133**	**96.8**	**97.5**	**0.165**	**0.133**
	QMR	**87.6**	**82.9**	**0.458**	**0.538**	**87.6**	**82.9**	**0.458**	**0.538**
	Empty	16.7	16.8	1.935	2.002	16.7	16.8	1.935	2.002
	Shuffled	38.1	39.9	2.916	3.146	38.1	39.9	2.916	3.146
Codex	GPT-5.6 Sol	64.6	49.3	0.925	1.385	64.9	50.8	0.926	1.323
	DeepSeek-V4-Flash	41.4	31.2	1.725	2.157	40.1	30.0	1.795	2.319
	GLM-5.2	40.9	28.7	1.657	2.046	42.5	28.1	1.622	2.046
Claude Code	Claude Sonnet 5	43.7	28.8	1.499	1.882	42.1	28.8	1.588	1.964
	DeepSeek-V4-Flash	49.2	34.0	1.641	2.027	45.3	31.8	1.723	2.190
	GLM-5.2	45.7	28.7	1.510	2.067	42.2	29.0	1.672	2.057
AI scientists	AI-Scientist-v2	59.2	43.8	1.031	1.448	-	-	-	-
	AutoDiscovery	27.5	27.1	1.474	1.526	-	-	-	-

Evaluated AI scientists and Agents. (1) **General Agents.** General AI agents are increasingly being used to assist and accelerate scientific discovery. We evaluate Codex (OpenAI, 2026a) with GPT-5.6 Sol (OpenAI, 2026b) and Claude Code (Anthropic, 2026a) with Claude Sonnet 5 (Anthropic, 2026b). To compare the two harnesses with shared model backends, we also pair each with DeepSeek-V4-Flash (DeepSeek-AI, 2026) and GLM-5.2 (GLM-5-Team et al., 2026). Each of the six harness-model configurations is evaluated with and without scientific skills. (2) **AI Scientists.** To test whether workflows designed for scientific discovery help with mechanistic hypothesis generation from observational data, we evaluate the ideation component of AI-Scientist-v2 (Yamada et al., 2025) and the Bayesian-surprise-driven search procedure of AutoDiscovery (Agarwal et al., 2026).

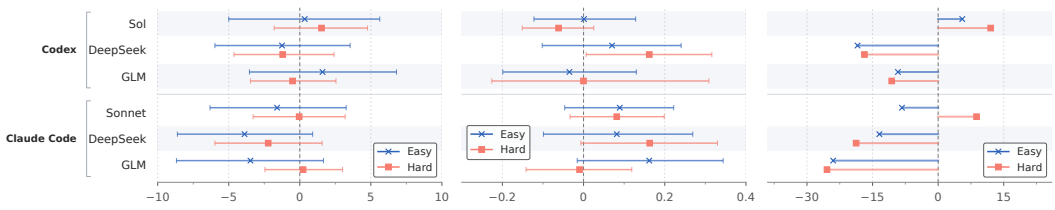
Implementation details. For skill-enabled runs, SkillRouter-Reranker-0.6B (Zheng et al., 2026) selects 32 skills from the collected pool (Section 3.3) using the task description and public dataset information, including its overview, structure, and format. General-agent runs use a 30-minute terminal budget and high reasoning settings. Across the twelve general-agent configurations, 214-289 of the 300 tasks yield valid terminal outputs. Invalid or missing outputs are scored using the result from the empty-control condition. Both AI-scientist systems use GPT-5.6 Sol. AutoDiscovery is evaluated on a fixed, stratified subset of 40 tasks, comprising 20 easy and 20 hard, since its search used 4,464 accepted LLM calls, averaging 111.6 per task. GPT-5.6 Terra (OpenAI, 2026c), with high reasoning and temperature zero, serves as the fixed hypothesis interpreter. Details of interpretation and scoring are given in Section 3.6.

4.2 Evaluating the Ability of Generating Data-Driven Mechanistic Hypothesis

Table 3 shows performance across the evaluated agents and AI Scientists on easy and hard tasks. Across both difficulty levels, agent-generated hypotheses achieve lower MCA and higher MCNLL than the QMR and PMR. Codex with GPT-5.6 Sol and scientific skills achieves the highest MCA across the evaluated configurations, but substantial gaps remain relative to the QMR (agent minus QMR, easy/hard: MCA $-22.7\%/-32.1\%$; MCNLL $+0.468/+0.785$). The gaps are even larger relative to the PMR. Both empty and shuffled controls also perform substantially worse than the MRs on both metrics, supporting the evaluation's validity to distinguish task-relevant mechanistic content from absent or mismatched content. These results indicate substantial room for improvement in AI agents' ability to generate mechanistic hypotheses from the available observations.

4.3 Evaluating Hypothesis Generation Quality and Efficiency across Harnesses and Skills

Harness choice affects performance even when the model backend and skill setting are held fixed (Figure 4). Claude Code generally achieves higher MCA than Codex with DeepSeek-V4-Flash and GLM-5.2. The clearest advantage appears with DeepSeek-V4-Flash, where both MCA and MCNLL favor Claude Code under both skill settings; without skills, the easy-task MCA difference is 7.8 percentage points (95% CI: [2.2, 13.3]). These gains accompany a different execution pattern: Claude Code makes more tool calls in all four matched comparisons (Appendix E.3 for details) and takes longer than Codex with DeepSeek-V4-Flash under both skill settings. With GLM-5.2, Claude Code is slightly slower without skills but faster with skills. Thus, the harness influences both hypothesis-generation performance and execution cost.



(a) MCA Changes (pp) (b) MCNLL Changes (c) Runtime Changes (%)

Figure 4: Changes associated with scientific-skill access. Points show with-skills minus without-skills differences for easy and hard tasks; horizontal bars denote 95% confidence intervals. Panels report changes in (a) MCA, (b) MCNLL, and (c) agent execution time.

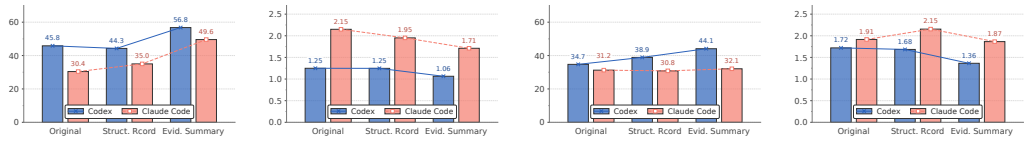
Access to scientific skills reduces execution time for DeepSeek-V4-Flash and GLM-5.2 under both harnesses, but these gains do not consistently translate into better mechanistic hypotheses. Mean execution time decreases by 10.1-18.1% across these four configurations. However, the pattern is model-dependent: Codex with GPT-5.6 Sol takes longer with skills, whereas Claude Code with Claude Sonnet 5 shows little runtime change. Nevertheless, MCA and MCNLL show limited or inconsistent improvements. Together, these results suggest that making scientific guidance available can improve execution efficiency, but does not consistently improve the mechanistic content of the resulting hypotheses.

Takeaways.
- *Harness choice affects both hypothesis quality and execution cost.*
- *Scientific skill access can shorten execution without consistently improving mechanistic hypothesis generation.*

4.4 Evaluating Mechanistic Hypothesis Generation in AI Scientists

AI-scientist workflows do not consistently improve mechanistic hypothesis generation. AI-Scientist-v2 achieves 59.2% and 43.8% MCA on easy and hard tasks, respectively, below Codex with GPT-5.6 Sol. AutoDiscovery performs even lower in the matched 40-task comparison, reaching 27.5% and 27.1% MCA, compared with 47.1% and 35.8% for Codex with GPT-5.6 Sol without skills. Their MCNLL profiles differ from their MCA rankings, with AutoDiscovery assigning higher probability to the ground-truth direction on the hard subset despite its lower MCA. These results show that incorporating scientific-discovery workflows changes performance profiles but does not by itself improve agents' ability to formulate mechanistic hypotheses from observations.

Takeaways.
- *AI-scientist workflows do not consistently improve mechanistic hypothesis generation over a general agent using the same underlying model.*



(a) Easy Task / MCA (b) Easy Task / MCNLL (c) Hard Task / MCA (d) Hard Task / MCNLL

Figure 5: Interventions on 40 tasks per configuration. Lines connect scores across conditions; missing outputs use the empty control.

4.5 Evaluating Effects of Evidence Access and Retention

We examine two settings that change how agents access and retain evidence from the public observations. In the **evidence summaries** setting, agents receive precomputed summaries of how responses vary across conditions in the public data. These summaries make relevant empirical relationships explicit without providing new observations. In the **structured records** setting, agents receive no precomputed evidence. Instead, they are instructed to record their own findings, the conditions under which each finding holds, its supporting evidence, and potential falsifiers, and to revisit these records before producing their final hypotheses. We evaluate both settings on fixed 40 tasks using Codex with GPT-5.6 Sol and Claude Code with DeepSeek-V4-Flash, the highest-MCA model configurations within their respective harnesses in the main experiments.

Providing evidence summaries improves MCA by 10.0 percentage points for both configurations, with MCNLL reductions of 0.288 for Codex with GPT-5.6 Sol and 0.244 for Claude Code with DeepSeek-V4-Flash (Figure 5). MCNLL also improves on tasks with accepted outputs in all settings. These results suggest that making empirical relationships explicit can help agents use the available observations to formulate mechanistic hypotheses. Nevertheless, performance remains below the MRs, indicating that supplying these findings does not eliminate the difficulty of inferring the generating process. Structured records yield smaller MCA gains of 1.9 and 2.1 percentage points for Codex with GPT-5.6 Sol and Claude Code with DeepSeek-V4-Flash, respectively, without consistent improvement in MCNLL. For Codex with GPT-5.6 Sol, the resulting hypotheses more accurately imply unchanged responses but less accurately imply increases or decreases.

These results suggest that the difficulty of mechanistic hypothesis generation for AI agents lies partly in discovering and selecting relevant evidence from the available observations. Structured records provide a way to retain findings during extended analysis, but cannot compensate for evidence that agents have overlooked. Their limited gains further suggest that preserving findings and integrating them into a mechanistic hypothesis are distinct demands; support for the former may not fully address the latter.

Takeaways.

- *Evidence summaries improve mechanistic hypothesis generation more consistently than structured records, suggesting that agents leave useful evidence in the available observations unexploited.*
- *The limited gains from structured records suggest that retaining findings and using them to formulate a mechanistic hypothesis are distinct demands.*

5 Conclusion

This paper introduces MechHypoBench, a benchmark for data-driven mechanistic hypothesis generation that combines paper-derived mechanisms across 14 fields with independently collected empirical datasets. We then develop an evaluation framework that assesses open-form mechanistic hypotheses through their consequences under withheld conditions, with ground truth outcomes determined by the same paper-derived mechanism that generates the observations. Through systematic experiments with AI agents, we identify limitations in data-driven mechanistic hypothesis generation and examine how evidence availability and the retention of findings affect the hypotheses they produce.

AI USE STATEMENT

In this work, we used generative AI tools to assist with extracting mechanistic claims from scientific papers, as described in Section 3.2, Section 3.4 and Appendix C.2. Additionally, we used generative AI tools for editing source codes, and polishing the manuscript. No other required-disclosure categories apply. We reviewed all AI-assisted outputs and take responsibility for the final content of this work, including the text and artifacts.

ETHICS STATEMENT

This work uses datasets obtained from public scientific repositories and does not involve new data collection or interaction with human participants. We use the data to construct benchmark tasks and do not make claims about the individuals, groups, or populations represented in the source studies. Source datasets are used and distributed only in accordance with their applicable licenses and access conditions; records that cannot be redistributed are identified by their original repository references. The benchmark generates task variables and response channels assigned experimental semantic roles from paper-derived mechanisms. These generated tasks should not be interpreted as clinical, behavioral, or scientific conclusions about the original source datasets. The benchmark is intended to evaluate scientific AI systems.

REPRODUCIBILITY STATEMENT

The appendix provides details of dataset collection, task construction, experimental settings, and supplementary result analyses. We release a task subset together with the evaluation framework in `https://anonymous.4open.science/r/ICLR-2027-benchmark-E1FB`. The remaining tasks, construction, and benchmark artifacts will be released after the review process.

REFERENCES

Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, et al. Autodiscovery: Open-ended scientific discovery via bayesian surprise. *Advances in Neural Information Processing Systems*, 38:25181–25219, 2026.

agentclientprotocol. claude-agent-acp. `https://github.com/agentclientprotocol/claude-agent-acp`, 2026a. Official software repository. Accessed September 26, 2026.

agentclientprotocol. codex-acp. `https://github.com/agentclientprotocol/codex-acp/`, 2026b. Official software repository. Accessed September 26, 2026.

Anthropic. Claude Code. `https://github.com/anthropics/claude-code`, 2026a. Official software repository. Accessed September 11, 2026.

Anthropic. Claude Sonnet 5. `https://platform.claude.com/docs/en/models/sonnet-5/overview`, 2026b. Official model documentation. Accessed September 11, 2026.

benchflow-ai. benchflow. `https://github.com/benchflow-ai/benchflow`, 2026. Official software repository. Accessed September 26, 2026.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, Sergey Feldman, Dany Haddad, Jena Hwang, Peter Jansen, Varsha Kishore, et al. Astabench: Rigorous benchmarking of ai agents with a scientific research suite. In *International conference on learning representations*, volume 2026, pp. 110136–110223, 2026.

Tingting Chen, Beibei Lin, Zifeng Yuan, Qiran Zou, Hongyu He, Anirudh Goyal, Yew-Soon Ong, and Dianbo Liu. Hypospace: A diagnostic benchmark for set-valued hypothesis generation under underdetermination and sublinear coverage bounds. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=QpjtK65JHO`.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pp. 96934–96990, 2025.

Carl F Craver. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon press, 2007.

Carl F Craver and Lindley Darden. *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press, 2013.

Lindley Darden. Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science*, 69(S3):S354–S365, 2002.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Janine Felden, Lars Möller, Uwe Schindler, Robert Huber, Stefanie Schumacher, Roland Koppe, Michael Diepenbroek, and Frank Oliver Glöckner. Pangaea-data publisher for earth & environmental science. *Scientific Data*, 10(1):347, 2023.

Erin D Foster and Ariel Deardorff. Open science framework (osf). *Journal of the Medical Library Association: JMLA*, 105(2):203, 2017.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, pp. 1–3, 2026.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M Williams, Stefan Bekiranov, and Aidong Zhang. Ideabench: Benchmarking large language models for research idea generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 5888–5899, 2025.

Lin He and Vinita Nahar. Reuse of scientific data in academic publications: An investigation of dryad digital repository. *Aslib Journal of Information Management*, 68(4):478–494, 2016.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Andrew J Lew, Yuan Cao, and Markus J Buehler. Projectionbench: Evaluating scientific hypothesis generation in llms under progressive information disclosure. *arXiv preprint arXiv:2605.30284*, 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.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13187–13207, 2026.

Qingqing Long, Haotian Chen, Chenyang Zhao, Xiaolei Du, Xuezhi Wang, Pengyao Wang, Chengzan Li, Yuanchun Zhou, and Hengshu Zhu. Sciencedb ai: An llm-driven agentic recommender system for large-scale scientific data sharing services. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 7715–7726, New York, NY, USA, 2026. ISBN 9798400722592.

Peter Machamer, Lindley Darden, and Carl F Craver. Thinking about mechanisms. *Philosophy of science*, 67(1):1–25, 2000.

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*, volume 2025, pp. 4556–4579, 2025.

Stephanie D Nofal, Caia Dominicus, Malgorzata Broncel, Nicholas J Katris, Helen R Flynn, Gustavo Arrizabalaga, Cyrille Y Botté, Brandon M Invergo, and Moritz Treeck. A positive feedback loop mediates crosstalk between calcium, cyclic nucleotide and lipid signalling in calcium-induced toxoplasma gondii egress. *PLoS Pathogens*, 18(10):e1010901, 2022.

Ethan F Oblak, Jarrod A Lewis-Peacock, and James S Sulzer. Self-regulation strategy, feedback timing and hemodynamic properties modulate learning in a simulated fmri neurofeedback environment. *PLoS computational biology*, 13(7):e1005681, 2017.

OpenAI. Codex CLI. `https://github.com/openai/codex`, 2026a. Official software repository. Accessed September 11, 2026.

OpenAI. GPT-5.6 Sol. `https://developers.openai.com/api/docs/models/gpt-5.6-sol`, 2026b. Official model documentation. Accessed September 11, 2026.

OpenAI. GPT-5.6 Terra. `https://developers.openai.com/api/docs/models/gpt-5.6-terra`, 2026c. Official model documentation. Accessed September 11, 2026.

Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

Parshin Shojaee, Ngoc-Hieu Nguyen, Kazem Meidani, Amir Barati Farimani, Khoa D Doan, and Chandan K Reddy. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*, 2025.

Miguel-Angel Sicilia, Elena García-Barriocanal, and Salvador Sánchez-Alonso. Community curation in open dataset repositories: insights from zenodo. *Procedia Computer Science*, 106:54–60, 2017.

Zhangde Song, Jieyu Lu, Yuanqi Du, Botao Yu, Thomas M Pruyn, Yue Huang, Kehan Guo, Xiuzhe Luo, Yuanhao Qu, Yi Qu, et al. Evaluating large language models in scientific discovery. *arXiv preprint arXiv:2512.15567*, 2025.

Juliette Tariel, Sandrine Plénet, and Émilien Luquet. Transgenerational plasticity of inducible defences: Combined effects of grand-parental, parental and current environments. *Ecology and Evolution*, 10(5):2367–2376, 2020.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023a.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023b.

James Woodward and James Francis Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

James F Woodward. Data and phenomena: A restatement and defense. *Synthese*, 182(1):165–179, 2011.

Guangzhi Xiong, Eric Xie, Corey Williams, Myles Kim, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models. *arXiv preprint arXiv:2505.14599*, 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.

YanZhao Zheng, ZhenTao Zhang, Chao Ma, YuanQiang Yu, JiHuai Zhu, Yong Wu, Tianze Xu, Baohua Dong, Hangcheng Zhu, Ruohui Huang, and Gang Yu. Skillrouter: Skill routing for llm agents at scale, 2026. URL https://arxiv.org/abs/2603.22455.

APPENDIX CONTENTS

A Limitations and Future Directions 15

B Comparison with Closely Related Benchmarks 15

C Scientific Sources and Benchmark Composition 15

C.1 Manual Review of Mechanistic Explanations in Empirical Research 15

C.2 Paper Collection and Claim Extraction 16

C.3 Dataset Preparation and Role Annotation 16

C.4 Benchmark Composition 17

D Task Construction and Evaluation Details 19

D.1 Paper-Derived Mechanisms 19

D.2 Dataset Alignment and Observation Generation 20

D.3 Matched Comparisons and Observation Withholding 21

D.4 Mechanism References (MRs) 23

D.5 Hypothesis Interpretation and Evaluation 24

E Additional Experimental Details and Results 25

E.1 Experimental Settings 25

E.2 Complete Results and Evaluation Controls 26

E.3 Agent Harnesses and Scientific Skills 27

E.4 AI Scientists 29

E.5 Evidence Summaries and Structured Records 29

E.6 Hypothesis Selection Diagnostic 31

F Case Studies 32

F.1 Successful Mechanistic Hypothesis 32

F.2 Missing a Conditional Relationship 33

F.3 Effects of Evidence Summaries and Structured Records 34

A Limitations and Future Directions

Transfer to empirical scientific discovery. Our benchmark combines empirical dataset structure with responses generated by paper-derived mechanisms. This construction supports controlled assessment while retaining observational complexity, but performance in this setting does not directly establish an agent's ability to discover mechanisms in real-world scientific research. Future work could extend evaluation to empirical studies in which proposed hypotheses are assessed through additional experiments.

Interactive scientific investigation. Agents formulate mechanistic hypotheses from a fixed set of public observations. Although they can perform iterative analyses, they cannot obtain new experimental measurements during hypothesis generation. Extending the benchmark to allow agents to propose experiments and receive additional observations would support evaluation of how they test and revise mechanistic hypotheses throughout the discovery process.

B Comparison with Closely Related Benchmarks

DiscoveryBench. DiscoveryBench evaluates discoveries through variables, contexts, and relationships supported by data (Majumder et al., 2025). Recovering these relationships does not by itself establish whether an agent can formulate a mechanistic hypothesis describing how interacting processes produce a phenomenon. Its real-data tasks draw on published findings and associated datasets without generally providing fully specified generating mechanisms, while its synthetic tasks offer controlled variation using generated data that may simplify empirical complexity. MechHypoBench combines paper-derived mechanisms with independently collected empirical datasets to assess mechanistic hypotheses against a known generating process while retaining empirical observational complexity. Its central target is the mechanistic hypothesis connecting individual findings into an account of how a phenomenon arises.

LLM-SRBench. LLM-SRBench evaluates scientific equation discovery using transformed physical equations and synthetic scientific models, assessing both numerical prediction and symbolic agreement with the generating equations (Shojaee et al., 2025). Its hypotheses must specify mathematical relationships that can be evaluated numerically. MechHypoBench accepts open-form mechanistic accounts of processes, interactions, and conditions of applicability, without requiring a complete mathematical specification. It also retains the heterogeneous observational structure of empirical scientific datasets, extending evaluation to agents' analysis of such data and the mechanistic hypotheses they produce.

HypoBench. HypoBench evaluates LLM's ability of natural-language hypothesis generation across real-world and synthetic settings (Liu et al., 2025). Its real-world tasks primarily concern classification of textual and behavioral outcomes, including deception, persuasion, mental stress, engagement, and citations. Its synthetic tasks likewise center on discovering features and relationships that support outcome prediction, with settings such as elections, college admissions, and consumer choices. MechHypoBench instead targets mechanistic accounts of scientific phenomena across 14 fields, assessing AI agents to explain how processes and interactions produce observed behavior.

C Scientific Sources and Benchmark Composition

C.1 Manual Review of Mechanistic Explanations in Empirical Research

Sample collection. We reviewed 490 abstracts published between 2021 and 2025, with 70 abstracts from each of seven journals: Nature, Science, Cell, PNAS, Nature Physics, Nature Chemistry, and Nature Geoscience. The sampling frame was constructed using OpenAlex source identifiers, restricted to articles with available abstracts and excluding retracted papers. No mechanism-related keyword was used for selection.

Review criteria. Reviewers applied written criteria to determine whether each abstract presented a mechanism as an explanatory output. An abstract qualified when it reported using data to propose or materially refine an explanatory process, or to test, discriminate between, or constrain existing mechanistic accounts. An exact supporting quotation was recorded for each positive judgment.

Results. As shown in Figure 1, 344 of the 490 abstracts (70.2%) presented a mechanism as an explanatory output. Of these, 291 proposed or materially refined a mechanism, while 53 tested an existing mechanism. These results describe the sampled abstracts, with equal representation from the seven journals.

C.2 Paper Collection and Claim Extraction

Collection sources. We collected candidate papers through Europe PMC, publisher interfaces for PLOS and eLife, and OpenAlex. Europe PMC supported collection in the life sciences and medicine, including targeted searches for studies involving prior exposure, carryover effects, and changes across process stages. Publisher interfaces provided structured full texts, while OpenAlex supported broader disciplinary coverage and identified open-access locations. Records were normalized and deduplicated by DOI. Table 4 summarizes the collection and retention stages.

Table 4: Paper collection and retention across construction stages. Counts refer to papers before mechanism construction and paper-derived mechanism assets thereafter.

Stage	Count
Deduplicated paper catalog	2,713
Within selected OpenAlex fields	2,659
Prioritized for full-text acquisition	1,950
Structured full texts obtained	1,465
Papers processed for mechanism construction	599
Accepted mechanism assets	335
Retained after licensing checks	306
Used in the final benchmark	244

Scientific review. Candidate papers were assessed for empirical grounding and claims explaining how a phenomenon arises. We excluded reviews, protocols, retracted papers, and studies concerned solely with prediction or optimization. Review examined whether the reported evidence supported the proposed processes and relationships, and whether the experimental semantic roles and their conditions of applicability could be identified. Eligible papers were prioritized for full-text acquisition with attention to disciplinary coverage.

Claim extraction and verification. LLMs assisted in identifying mechanistic claims and their supporting passages. Each extracted claim retained its location in the source article and the conditions under which it applied. Reviewers checked the extracted content against the original text, including the proposed experimental roles and the grouping of related claims. Additional claims introduced during review also required explicit source support. These reviewed claims provided the scientific basis for the paper-derived mechanisms described in Section 3.4.

Retained collection. Mechanism construction was performed on a subset of 599 of the 1,465 available full texts; the remaining papers were not processed at this stage. Of the 335 accepted mechanism assets, 306 remained after licensing checks, spanning 14 scientific fields. The final benchmark uses 244 of these assets to construct 300 tasks. Retention in the mechanism collection precedes task construction: an asset contributes to the final benchmark when it can be paired with an empirical dataset to produce suitable matched comparisons and is selected for the evaluated task set.

C.3 Dataset Preparation and Role Annotation

Dataset collection. We independently collected scientific datasets from ScienceDB, PANGAEA, Zenodo, the Open Science Framework, and Dryad. Candidate datasets were assessed for traceable provenance, accessible files, suitable reuse licenses, and documentation describing the measured variables. Collection combined repository searches with inspection of previously acquired records. Duplicate versions and overlapping subsets of the same source dataset were consolidated before task construction. Table 5 summarizes the repository distribution of the prepared collection and the datasets used in the final benchmark.

Table 5: Dataset sources in the prepared collection and final benchmark. Each dataset is counted once within each column.

Dataset Repository Source	# Prepared collection	# Used in final tasks
ScienceDB (Long et al., 2026)	87	77
PANGAEA (Felden et al., 2023)	57	35
Zenodo (Sicilia et al., 2017)	42	27
Open Science Framework (Foster & Deardorff, 2017)	10	8
Dryad He & Nahar (2016)	4	3
Total	200	150

Data preparation. We inspected source files to identify their formats, table organization, variable types, missing values, and repeated observations. For datasets containing multiple tables, preparation selected an appropriate table or applied documented joins and reshaping operations. Dataset-specific preprocessing accommodated differences in file formats and data structures. The prepared data retained links to the original measurements, together with grouping and ordering information needed to interpret the observations.

Experimental role annotation. Reviewers used dataset documentation and variable descriptions to annotate the experimental semantic roles defined in Section 3.1. Table 6 summarizes the source information supporting each assignment. Role annotation preserved the scientific meaning of the recorded variables: temporal or process ordering required support from the study design, and history required information about antecedent exposure or state. Subject, site, batch, or replicate identifiers were retained when available, together with auxiliary measurements. The assignment of these roles was completed before pairing datasets with paper-derived mechanisms.

Table 6: Source information used to annotate experimental semantic roles.

Role	Supporting source information
Driver	A documented factor or exposure whose effect can be examined across different values or levels.
Context	A distinct environmental or system condition that may modify the driver's effect.
History	Information about antecedent exposure or state, supported by the study design or documented ordering of observations.
Phase	A documented temporal or process coordinate identifying the stage at which observations were collected.

Retained datasets. Dataset retention required executable preparation, scientifically supported role assignments, and sufficient variation and observational structure for task construction. Datasets were excluded when their documentation did not support the required roles or when preparation and reuse requirements could not be met. The resulting collection contains 200 prepared datasets with reviewed role annotations. Of these, 150 contributed to the final 300 tasks following mechanism-dataset pairing and task selection. Dataset sizes, observational structures, and their distribution across the benchmark are reported in the benchmark composition analysis.

C.4 Benchmark Composition

The benchmark comprises 300 tasks, including 120 easy and 180 hard tasks, constructed from mechanisms derived from 244 papers and observations from 150 independently collected scientific datasets.

Scientific coverage. Table 7 reports the distribution across 14 scientific fields. Each source paper is counted once in the unique-paper column, while task counts include repeated use of a paper-derived mechanism with different datasets. All fields are represented in both difficulty groups, although their frequencies differ.

Table 7: Scientific-field coverage of the final benchmark. Fields follow the primary OpenAlex classification of the source papers.

Field	# Unique Papers	# Easy	# Hard	# Total tasks
Agricultural and Biological Sciences	35	12	27	39
Biochemistry, Genetics and Molecular Biology	36	12	36	48
Chemical Engineering	8	3	6	9
Chemistry	2	2	1	3
Earth and Planetary Sciences	14	6	10	16
Energy	11	5	7	12
Engineering	15	9	11	20
Environmental Science	18	9	13	22
Immunology and Microbiology	23	17	13	30
Materials Science	8	4	8	12
Medicine	22	11	16	27
Neuroscience	36	22	23	45
Pharmacology, Toxicology and Pharmaceutics	5	4	2	6
Physics and Astronomy	11	4	7	11
Total	244	120	180	300

Dataset scale and observational complexity. The 150 unique source datasets collectively contain 17,978,782 observation records. Table 8 summarizes their sizes and dimensionality. These statistics count each source dataset once, regardless of task reuse, and describe the source data before task-specific observation withholding.

Table 8: Size and dimensionality of the 150 unique source datasets.

Measure	Minimum	Median	Maximum
Observation records	262	5,131	5,996,958
Columns	6	25	588

Table 9: Observational designs of the 150 unique source datasets.

Observational design	Datasets
Spatiotemporal	72
Longitudinal	43
Spatial cross-sectional	14
Grouped	13
Time-series/progression	6
Cross-sectional/ordered	2
Total	150

The datasets span six observational designs (Table 9). Spatiotemporal and longitudinal or repeated-measurement datasets account for the largest shares, while spatial cross-sectional, grouped, time-series, and ordered cross-sectional datasets provide additional structures. Explicit grouping information, such as subjects, sites, or batches, is available in 121 datasets. Missingness also varies: 49 datasets contain no missing values, 61 have a missing fraction above zero and at most 10%, 20 have a missing fraction above 10% and at most 30%, and 20 exceed 30%. Across datasets, the median missing fraction is 1.54%. These properties require agents to handle variation in scale, dimensionality, grouping, ordering, and data completeness when analyzing the public observations.

D Task Construction and Evaluation Details

D.1 Paper-Derived Mechanisms

LLM-assisted construction and human review. We construct paper-derived mechanisms from the reviewed claims and supporting passages described in Appendix C.2. LLMs assist in organizing these claims into a directed graph of processes, interactions, and responses. Each graph component retains links to its supporting claims and their conditions of applicability. Reviewers assess whether the constructed mechanism preserves the scientific meaning of the reported processes and relationships. Human review provides the scientific assessment, while automated checks verify structural and numerical requirements.

Numerical realization. The graph-generation stage specifies both the dependencies and the numerical functions and parameters assigned to its nodes. Inputs to a node are experimental-role values or outputs of upstream nodes. The implementation supports three operations:

- **Weighted contributions and interactions (`linear_terms`).**

$$u_v = b_v + \sum_j w_{vj} \prod_{x \in A_{vj}} x,$$

  where u_v is the output of node v, b_v is its constant term, and A_{vj} contains the inputs to term j. A single input represents an individual contribution, while multiple inputs express an interaction. Repeated factors allow powers of an input.
- **Saturating transformation.**

$$u_v = \frac{1}{1 + \exp\big[-\operatorname{clip}\big(s_v(x - a_v), -20, 20\big)\big]},$$

  where x is the input, $s_v \neq 0$ controls the direction and steepness of the transition, and a_v specifies its center. Clipping bounds the exponent during numerical evaluation.
- **Exponential decay (`exp_decay`).**

$$u_v = \exp(-\lambda_v x),$$

  where $\lambda_v > 0$ controls the decay rate.

These operations provide an executable realization of the paper-derived relationships. Their parameters instantiate qualitative claims on normalized coordinates rather than estimate effect sizes from the original study. Deterministic assembly orders the graph by dependency and connects designated node outputs to response variables. Dataset-specific transformations, group-specific offsets, and measurement noise are introduced during observation generation and do not enter the mechanism's internal dependencies.

Validation and freezing. Structural validation checks the allowed operations and parameter constraints, dependency ordering, links to claim components, and response mappings. Graphs with cycles, undefined inputs, missing component coverage, or invalid response mappings are rejected. Behavioral validation evaluates the numerical mechanism on a fixed grid of conditions before any dataset is paired with it. These checks examine whether the represented interactions and nonlinear operations affect the responses, whether response channels capture distinct behavior, and whether the mechanism supports condition-dependent changes and structurally absent effects. Passing these checks establishes compliance with the construction requirements; scientific consistency is assessed through human review. Mechanisms requiring scientific revision are reconstructed and checked again. The accepted structure, functions, and parameters are fixed as $\mathcal{M}^*$ before final dataset pairing. Algorithm 1 summarizes the procedure.

In Algorithm 1, G includes the graph structure, node functions, numerical parameters, response mappings, and links to supporting claims. N denotes its executable numerical representation, and U contains the node outputs evaluated over the canonical condition grid $\mathcal{C}_{\text{grid}}$. Structural and behavioral checks follow the requirements described above. The human-review decision is $d \in \{\text{accept}, \text{revise}, \text{reject}\}$; a mechanism returned for revision is reconstructed and submitted to the same checks again.

Algorithm 1 Construction of a paper-derived mechanism

Require: Reviewed claims E, supporting passages P, role annotations A, and canonical condition grid $\mathcal{C}_{\mathrm{grid}}$.
Ensure: Accepted mechanism $\mathcal{M}^*$, rejection, or a request for revision.

$G \leftarrow \mathrm{ConstructGraph}_{\mathrm{LLM}}(E, P, A)$
if $\neg\, \mathrm{ValidStructure}(G, E)$ **then**
  return structural rejection
end if
$N \leftarrow \mathrm{AssembleNumericalMechanism}(G, E)$
$U \leftarrow \mathrm{Evaluate}(N, \mathcal{C}_{\mathrm{grid}})$
if $\neg\, \mathrm{ValidBehavior}(N, U)$ **then**
  return behavioral rejection
end if
$d \leftarrow \mathrm{HumanReview}(E, P, A, G, U)$
if $d = \mathrm{revise}$ **then**
  return revision required
else if $d = \mathrm{reject}$ **then**
  return scientific rejection
end if
$\mathcal{M}^* \leftarrow \mathrm{Freeze}(N)$
return $\mathcal{M}^*$

D.2 Dataset Alignment and Observation Generation

Dataset alignment. We use the reviewed experimental-role annotations described in Appendix C.3 to connect an independently collected scientific dataset to the fixed paper-derived mechanism $\mathcal{M}^*$. Dataset-specific preprocessing loads the annotated variables and performs any documented table joins or reshaping operations. The resulting data provide the conditions under which the mechanism is evaluated, while retaining source measurements, auxiliary variables, and available grouping and ordering information.

Condition construction. Variables assigned to the driver and context roles are converted to percentile ranks, producing normalized coordinates in $[0, 1]$. The variable associated with history is normalized in the same way before its preceding values are aggregated. Missing values are median-filled for these internal computations; this does not replace missing entries in the retained source columns.

Observations are ordered within groups using the annotated temporal or process coordinate. Documented subject, site, batch, or replicate identifiers define groups when available. Otherwise, the implementation constructs groups from calendar years or quantiles of the ordering coordinate. These derived groups organize the numerical construction and need not correspond to original experimental units. The history coordinate is a rolling mean of preceding normalized history values within each group. The first observation, for which no preceding value is available, uses its own normalized history value. Phase is represented by the within-group percentile rank of the temporal or process coordinate. These transformations produce the condition vector $c = (c_{\mathrm{d}}, c_{\mathrm{x}}, c_{\mathrm{h}}, c_{\mathrm{p}})$ supplied to $\mathcal{M}^*$.

Observation generation. We generate the responses $R_{\mathcal{M}^*}(c)$ introduced in Section 3.1 by applying the fixed mechanism to the constructed conditions. Its designated node outputs are rescaled to obtain noise-free responses, denoted by μ_{ik} for row i and response channel k. The observation layer adds small perturbations to represent measurement variability and group-level differences. These include group-specific offsets shared by observations in the same group, variation shared across response channels, and channel-specific noise. Auxiliary measurements and missingness determine part of the observational variation, while random components are generated using fixed seeds.

The resulting recorded responses y_{ik} are clipped to $[0, 100]$ and retained alongside source measurements and auxiliary information. Each y_{ik} is a recorded realization of response channel k in $R_{\mathcal{M}^*}(c_i)$; repeated observations under the same condition may differ because of observational variation. The same observation-layer parameter settings are used across easy and hard tasks, without increasing noise specifically for hard tasks. Ground-truth direction of response changes are computed from the privately retained noise-free responses μ_{ik}, so the ground truth reflect the specified mechanism rather than a particular realization of measurement noise. Exact normalization, window, scaling, and noise parameters are specified in the implementation.

Algorithm 2 summarizes the construction. Here, g_i denotes the group assigned to row i, τ_i its ordering coordinate, and $C_D = (c_i)_i$ the row-wise condition vectors. A_{d}, A_{x}, and A_{h} identify the annotated source columns; h_0 contains normalized history values before aggregation. Percentile ranks use median-filled values internally. For an ordered group, $\mathrm{PrecedingMean}$ applies the rolling-history rule described above, including the first-observation fallback. The perturbation η_{ik} combines a group-specific offset, row-level variation shared across channels, and channel-specific noise; it is constructed jointly to preserve these dependencies.

The operation $\mathrm{ObservationVariation}$ uses A to identify the auxiliary columns in X, g to share offsets within groups, and the response-channel definitions and observation-layer coefficients stored with $\mathcal{M}^*$ to construct the perturbations. The seed s controls the random sampling.

Algorithm 2 Align a dataset and generate observations

Require: Fixed mechanism $\mathcal{M}^*$, scientific dataset D, reviewed role annotations A, and random seed s.
Ensure: Recorded observations Y, private noise-free responses μ, and corresponding conditions C_D.

$X \leftarrow \mathrm{PrepareDataset}(D, A)$
$(g, \tau) \leftarrow \mathrm{GroupingAndOrder}(X, A)$
for all $a \in \{\mathrm{d}, \mathrm{x}\}$ **do**
  $c_a \leftarrow \mathrm{PercentileRank}(X[A_a])$
end for
$h_0 \leftarrow \mathrm{PercentileRank}(X[A_{\mathrm{h}}])$
for all groups j **do**
  $I_j \leftarrow \mathrm{SortBy}_\tau(\{i : g_i = j\})$
  $c_{\mathrm{h}}[I_j] \leftarrow \mathrm{PrecedingMean}(h_0[I_j])$
  $c_{\mathrm{p}}[I_j] \leftarrow \mathrm{PercentileRank}(\tau[I_j])$
end for
$\eta \leftarrow \mathrm{ObservationVariation}(X, A, g, \mathcal{M}^*, s)$
for all rows i in X **do**
  $c_i \leftarrow (c_{\mathrm{d},i}, c_{\mathrm{x},i}, c_{\mathrm{h},i}, c_{\mathrm{p},i})$
  $V_i \leftarrow \mathrm{EvaluateNodes}(\mathcal{M}^*, c_i)$
  for all response channels k in $\mathcal{M}^*$ **do**
    $u_{ik} \leftarrow V_i[\mathrm{SourceNode}(\mathcal{M}^*, k)]$
    $\mu_{ik} \leftarrow \mathrm{ScaleResponse}(u_{ik})$
    $y_{ik} \leftarrow \mathrm{clip}_{[0,100]}(\mu_{ik} + \eta_{ik})$
  end for
end for
$C_D \leftarrow (c_i)_i$, $\mu \leftarrow (\mu_{ik})_{i,k}$, $Y \leftarrow (y_{ik})_{i,k}$
Retain Y alongside source measurements and auxiliary information; store μ privately.
return Y, μ, C_D

D.3 Matched Comparisons and Observation Withholding

Comparison construction. We construct matched comparisons from the observations generated in Appendix D.2. For this purpose, normalized driver, context, and history coordinates are grouped into low, middle, and high levels, while phase is divided into four ordered stages. A complete combination of these levels defines a condition group containing repeated observations. Each comparison $q = (c^{(1)}, c^{(0)}, R)$ changes one role level while holding the other three fixed. Thus, the comparison operates between groups of observations rather than isolated rows with exactly matching continuous coordinates. For each candidate comparison, the ground-truth direction of response change $\Delta_{\mathcal{M}^*}(q)$ is determined from the difference between the mean noise-free responses under the two conditions. Directional comparisons require a sufficiently large difference relative to observational variation. A no-change comparison additionally requires that the mechanism contain no directed dependency from the changed role to the examined response. Near-zero differences caused solely by cancellation are therefore insufficient to define a no-change ground-truth direction.

Robustness and mechanistic relevance. Candidate comparisons undergo three complementary stability checks: **(1) Observational stability.** Both condition groups must contain sufficient observations. For directional comparisons, the difference between recorded mean responses must agree with the noise-free direction and exceed a minimum magnitude. An additional check uses the estimated standard error of the difference to exclude comparisons whose assigned direction is unstable under measurement variation. **(2) Parameter robustness.** We evaluate 32 numerical variants

that preserve the mechanism's graph structure, coefficient signs, and qualitative transition bands. A comparison must retain its assigned direction and required effect magnitude in at least 90% of these variants, together with a nonnegative robustness margin. For no-change comparisons, the perturbed effects must remain within the specified equivalence range. **(3) Qualitative consistency.** Directional comparisons are further examined across 64 numerical realizations preserving the mechanism's qualitative properties. At least half must yield a directional change, and at least two-thirds of those directional changes must agree with the ground-truth direction. This reduces dependence on an exact private parameterization. These checks establish stability within the tested variations. Separately, we assess mechanistic relevance by comparing the full mechanism with simplified variants that remove interactions, linearize nonlinear transformations, or retain additive effects. A directional comparison is retained only when its ground-truth direction agrees with the full mechanism on the canonical condition grid and changes under at least one simplification. This ensures that the selected comparisons test conditional structure represented by the mechanism.

Each task contains twelve comparisons: five decreases, five increases, and two no-change cases. Selection distributes comparisons across experimental roles and response channels while maintaining the required stability and mechanistic-relevance checks.

Public and hidden observations. All observations belonging to either condition group of a selected comparison are withheld from the agent. Additional withholding removes selected groups and trailing observations within groups when sufficient data are available. The remaining rows form the public observations. Individual role levels may still occur publicly, but the complete combinations used in the hidden comparisons do not. The split must retain sufficient observations in both portions. Reference response changes are computed from the noise-free responses; recorded hidden observations are retained separately.

Difficulty assignment. After withholding, we apply a marginal baseline using only the public observations. For each directional comparison, this baseline compares the mean response at the two levels of the changed role without conditioning on the other roles. Its disagreements with the mechanism-derived reference indicate where marginal relationships fail to capture conditional behavior. Tasks with one to three disagreements across the ten directional comparisons are classified as easy, and those with six to ten as hard. Tasks with four or five disagreements are excluded from the evaluated benchmark. If an initial comparison set yields no disagreements, the constructor makes one constrained reselection attempt and recomputes the public split and baseline results. The replacement must satisfy the comparison-quality requirements and produce at least one disagreement; otherwise, the task is rejected. Difficulty is assigned before agent evaluation and does not depend on agent-generated hypotheses.

Algorithm 3 Construct matched comparisons and withhold observations

Require: Mechanism $\mathcal{M}^*$, recorded responses Y, noise-free responses μ, and row-wise conditions C_D.
Ensure: Public observations, hidden observations, comparison set $\mathcal{Q}$, reference changes, and difficulty.

$\mathcal{G} \leftarrow \mathrm{ConditionGroups}(C_D)$
$\mathcal{A} \leftarrow \emptyset$
for all pairs (g_1, g_0) in $\mathcal{G}$ differing in one role level **do**
  for all response channels R **do**
    $q \leftarrow (g_1, g_0, R)$
    $\delta_q \leftarrow \mathrm{ReferenceChange}(\mathcal{M}^*, \mu, q)$
    if $\mathrm{Stable}(q, \delta_q, Y, \mu, \mathcal{M}^*)$ and $\mathrm{Relevant}(q, \mathcal{M}^*)$ **then**
      $\mathcal{A} \leftarrow \mathcal{A} \cup \{(q, \delta_q)\}$
    end if
  end for
end for
$\mathcal{Q} \leftarrow \mathrm{SelectComparisons}(\mathcal{A}; 5, 2, 5)$
$(O_{\mathrm{public}}, O_{\mathrm{hidden}}) \leftarrow \mathrm{Withhold}(Y, C_D, \mathcal{Q})$
$d \leftarrow \mathrm{MarginalDisagreements}(O_{\mathrm{public}}, \mathcal{Q})$
if $d = 0$ **then**
  Attempt one constrained reselection from $\mathcal{A}$.
  Recompute the observation split and d.
end if
if selection or split checks fail, or $d = 0$ **then**
  return rejection
else if $1 \leq d \leq 3$ **then**
  $\mathrm{difficulty} \leftarrow \mathrm{easy}$
else if $6 \leq d \leq 10$ **then**
  $\mathrm{difficulty} \leftarrow \mathrm{hard}$
else
  return excluded transition task
end if
return $O_{\mathrm{public}}, O_{\mathrm{hidden}}, \mathcal{Q}, \{\delta_q : q \in \mathcal{Q}\}, \mathrm{difficulty}$

Here, Stable applies the observational and parameter checks, including qualitative consistency for directional comparisons. $\mathrm{Relevant}$ checks sensitivity to mechanism simplification for directional comparisons and the absence of a directed dependency for no-change comparisons. The selection counts $5, 2, 5$ correspond to decrease, no change, and increase, respectively.

D.4 Mechanism References (MRs)

MR construction. We construct the QMRs and PMRs from the same fixed paper-derived mechanism $\mathcal{M}^*$. Both use experimental-role and response names aligned with the task descriptions. The QMR is assembled from the mechanism's dependencies and semantic properties, with LLM assistance used to express the specified facts in text. The PMR is assembled from the same structure and its numerical specification. Checks verify the coverage of required facts, consistency of symbols and dependencies, and correspondence between the numerical expressions and the fixed mechanism.

MR content. The QMR describes how processes and interactions produce responses, including contribution signs, monotonicity, conditions of applicability, and qualitative transition bands. It may include symbolic expressions, but omits exact numerical coefficients and parameters. The PMR additionally provides the numerical functions, parameter values, and normalized condition coordinates needed to compute the mechanism's behavior. Table 10 summarizes their contents.

Table 10: Information provided by the two MRs. Both are derived from the same fixed mechanism.

Information	QMR	PMR
Experimental roles	Task-aligned role descriptions	Same descriptions with normalized coordinates
Dependencies	Processes, interactions, and response mappings	Same dependencies and mappings
Functional relationships	Contribution signs, monotonicity, and transition bands	Numerical functions and parameter values
Observation layer	Description of measurement variation and group-specific offsets	Corresponding numerical specification
Hidden response values	Not supplied	Not supplied
Matched-comparison answers	Not supplied	Not supplied

The numerical information in the PMR specifies the generating process; it does not contain computed hidden response values or answers to the matched comparisons. Comparison descriptions are supplied separately during evaluation, and the hypothesis interpreter must derive the implied response changes from the MR content.

Role in evaluation. Both MRs are processed by the same hypothesis interpreter used for agent-generated hypotheses. Their scores characterize performance when the relevant mechanism content is supplied at qualitative or numerical levels, providing context for interpreting the scores of generated hypotheses. Differences from these MRs indicate a gap relative to the supplied mechanism descriptions under the shared interpretation procedure. The MRs are not guaranteed performance upper bounds, and agent-generated hypotheses may themselves contain numerical parameters.

D.5 Hypothesis Interpretation and Evaluation

Hypothesis interpretation. Open-form mechanistic hypotheses need not provide a complete executable model, so their implications must be interpreted before comparison with ground truth outcomes. Asking the generating agent to perform this inference would combine the content of its hypothesis with its own inference capabilities and access to findings retained from the analysis. We therefore use a fixed LLM-based hypothesis interpreter to assess the content communicated in the submitted hypothesis under a common interpretation procedure. The interpreter receives one standalone mechanistic hypothesis together with the descriptions of the matched comparisons for that task. It has no access to public observations, hidden response values, reference labels, or other generated hypotheses. Withholding the observations prevents it from obtaining answers through a separate analysis of the data. It applies the supplied hypothesis without executing code, searching for external information, or conducting additional data analysis, and returns a probability distribution $p_h(t \mid q)$ over decrease, no change, and increase for each comparison q. The probabilities must sum to one. All generated hypotheses, MRs, and evaluation controls use the same interpreter configuration specified in Section 4.1.

Evaluation controls. The empty control supplies no hypothesis, measuring the interpreter's behavior when mechanistic content is absent. The shuffled control replaces the task's mechanistic result with a PMR from a different mechanism, while retaining the original comparison descriptions. Results are assigned across mechanisms with approximate length matching to limit differences attributable to text length. This control supplies detailed but mismatched mechanistic content. Both controls are evaluated against the original task's ground-truth response changes.

Scoring and aggregation. When an agent submits multiple hypotheses, the primary evaluation uses the hypothesis with the highest agent-assigned weight, as specified in the task instructions. Selection uses no hidden comparison outcomes. A deterministic evaluator computes MCA from the interpreter's most probable direction and MCNLL from the probability assigned to the mechanism-derived ground-truth direction, following Section 3.6. Probabilities are lower-bounded at 10^{-12} when computing logarithms. Each task score averages over its twelve matched comparisons, and reported performance averages task scores separately for the easy and hard groups. Invalid or missing

agent outputs are assigned the corresponding task's empty-control result, retaining those tasks in the reported averages. This keeps the evaluated task set fixed across agent configurations. Analyses restricted to tasks with valid outputs are identified separately and use the same task subset for each compared setting.

Uncertainty estimation. We estimate 95% confidence intervals using 5,000 bootstrap resamples of tasks. Each resampled task retains all its matched comparisons, preserving their within-task dependence. For paired comparisons, the same task indices are resampled jointly across settings before computing the score difference.

E Additional Experimental Details and Results

E.1 Experimental Settings

Execution settings. Each general-agent task runs in a fresh sandbox with access to the public task descriptions and data files. Hidden observations, matched comparisons, and MRs are unavailable during hypothesis generation, and task-runtime networking is disabled. We use `BenchFlow` (benchflow-ai, 2026) v0.7.0 to execute the general-agent configurations listed in Table 11. Codex CLI 0.147.0 is accessed through `codex-acp` (agentclientprotocol, 2026b) 1.4.0, and Claude Code 2.1.232 through `claude-agent-acp` (agentclientprotocol, 2026a) 0.69.0.

Table 11: General-agent configurations.

Harness	Version	Model
Codex	0.147.0	GPT-5.6 Sol
Codex	0.147.0	DeepSeek-V4-Flash
Codex	0.147.0	GLM-5.2
Claude Code	2.1.232	Claude Sonnet 5
Claude Code	2.1.232	DeepSeek-V4-Flash
Claude Code	2.1.232	GLM-5.2

For each task, SkillRouter-Reranker-0.6B selects 32 skills using the public task description and dataset information. In skill-enabled runs, the selected skills are placed in the designated skill directory of each agent harness, making them available through its skill-discovery and loading mechanism. Agents decide which skills to consult during analysis. Runs without skills omit the task-specific skill bundle while retaining the same task inputs, model backend, and execution budget.

General-agent runs receive a 30-minute execution budget per task, providing a common time allowance for iterative analysis while keeping the evaluation computationally manageable. Across the twelve configurations, valid outputs were obtained for 214-289 of the 300 tasks (71.3-96.3%). Thus, each configuration produced valid outputs for a majority of tasks within the allotted time; output validity is separate from the scientific quality assessed by MCA and MCNLL. Both AI-scientist systems use GPT-5.6 Sol. AI-Scientist-v2's ideation component is adapted to analyze the public files and produce mechanistic hypotheses. AutoDiscovery receives the public task descriptions and datasets through its discovery interface. Their outputs are converted to the common hypothesis format before evaluation. Runs affected by infrastructure or service errors were retried. These errors were recorded separately from invalid agent outputs and failures to finish within the execution budget. Valid outputs produced before a subsequent verification-system failure remained eligible for scoring when the corresponding interpretation and evaluation records were available.

Task subsets. General agents and AI-Scientist-v2 are both evaluated on the full 300 tasks (120 easy and 180 hard). AutoDiscovery is evaluated on a fixed subset selected by stratified sampling, comprising 20 easy and 20 hard tasks, because its search requires substantially more LLM calls and execution time per task. It used 4,464 accepted LLM calls, averaging 111.6 per task, compared with approximately 19.0 for AI-Scientist-v2. Despite this computational investment, AutoDiscovery achieves lower MCA than Codex with GPT-5.6 Sol on the same tasks, suggesting that extensive hypothesis search does not necessarily yield more accurate mechanistic hypotheses. Comparisons with other systems on this subset use the same task identifiers.

Hypothesis output. Agents may return one to five distinct, falsifiable mechanistic hypotheses, choosing how many to propose based on their analysis. Each hypothesis must provide a standalone account of the observed phenomenon. Scientific content is separated from supporting metadata, including confidence, weights, and evidence links. The primary evaluation uses the highest-weight hypothesis, as explicitly stated in the agent instructions, following Appendix D.5. Each hypothesis has an individual word limit to accommodate differences in mechanism complexity while bounding the length presented to the hypothesis interpreter. Each hypothesis is allowed up to 120% of the length of the task's PMR: $L_{\max} = \lfloor 1.2 L_{\text{ref}} \rfloor$, where L_{ref} is the PMR's whitespace-delimited word count. The PMR provides a task-specific length anchor because it describes the specified mechanism, and the additional allowance permits alternative wording and organization. The limit applies separately to each hypothesis, with no minimum length or shared limit across candidates. Agents receive the word limit and PMR length, but not the PMR content.

E.2 Complete Results and Evaluation Controls

Complete results. Tables 12 report MCA and MCNLL with 95% task-bootstrap confidence intervals for the evaluated agents, AI scientists, controls, and MRs. General agents and AI-Scientist-v2 are evaluated on all 120 easy and 180 hard tasks. AutoDiscovery is evaluated on 20 tasks from each difficulty group; comparisons restricted to the same tasks are presented in the supplementary AI-scientist analysis. For generated hypotheses, the primary score uses the highest-weight hypothesis. Invalid or missing outputs receive the corresponding empty-control result, following Appendix D.5. The valid/evaluated column reports output availability, not scientific correctness. Claude Code with GLM-5.2 without skills produces 222 valid outputs: 93 on easy tasks and 129 on hard tasks.

Evaluation controls. Both controls perform below the QMR and PMR on MCA and MCNLL. The shuffled control achieves higher MCA than the empty control but worse MCNLL, indicating that mismatched mechanistic content can yield more correct categorical decisions while assigning less probability to the correct outcomes overall. Several agent configurations also fall below the shuffled control on MCA while achieving lower MCNLL. These differences illustrate why both metrics are needed to characterize hypothesis performance.

Table 12: Easy- and hard-task results. Generated outputs use the highest-weight hypothesis, and missing or invalid outputs use the corresponding empty-control results. Brackets denote 95% task-bootstrap confidence intervals. AutoDiscovery† uses a 20-task subset for each difficulty. All other rows use 120 easy tasks and 180 hard tasks.

Model / MRs	Skills	Valid Rate (%) ↑		MCA (%)↑		MCNLL↓	
		Easy	Hard	Easy	Hard	Easy	Hard
Codex							
GPT-5.6 Sol	✗	96.7	96.1	64.6 [60.3, 68.9]	49.3 [45.7, 52.8]	0.925 [0.830, 1.022]	1.385 [1.290, 1.484]
GPT-5.6 Sol	✓	94.2	96.7	64.9 [60.4, 69.3]	50.8 [47.3, 54.3]	0.926 [0.820, 1.043]	1.323 [1.228, 1.421]
DeepSeek-V4-Flash	✗	73.3	76.7	41.4 [37.2, 45.6]	31.2 [28.4, 34.0]	1.725 [1.582, 1.872]	2.157 [1.957, 2.438]
DeepSeek-V4-Flash	✓	75.8	79.4	40.1 [36.0, 44.4]	30.0 [27.7, 32.4]	1.795 [1.649, 1.948]	2.319 [2.111, 2.607]
GLM-5.2	✗	74.2	69.4	40.9 [36.6, 45.3]	28.7 [26.3, 31.0]	1.657 [1.522, 1.796]	2.046 [1.943, 2.153]
GLM-5.2	✓	78.3	70.6	42.5 [38.3, 46.6]	28.1 [25.8, 30.6]	1.622 [1.495, 1.760]	2.046 [1.852, 2.328]
Claude Code							
Claude Sonnet 5	✗	80.8	85.0	43.7 [39.7, 47.7]	28.8 [26.3, 31.3]	1.499 [1.378, 1.618]	1.882 [1.798, 1.965]
Claude Sonnet 5	✓	84.2	83.9	42.1 [37.9, 46.4]	28.8 [26.2, 31.6]	1.588 [1.457, 1.716]	1.964 [1.860, 2.067]
DeepSeek-V4-Flash	✗	92.5	90.0	49.2 [45.1, 53.1]	34.0 [31.1, 37.1]	1.641 [1.490, 1.793]	2.027 [1.911, 2.145]
DeepSeek-V4-Flash	✓	95.0	91.7	45.3 [41.5, 49.2]	31.8 [29.1, 34.8]	1.723 [1.591, 1.857]	2.190 [2.069, 2.308]
GLM-5.2	✗	77.5	71.7	45.7 [41.4, 50.1]	28.7 [26.6, 31.0]	1.510 [1.366, 1.656]	2.067 [1.963, 2.178]
GLM-5.2	✓	79.2	80.6	42.2 [38.3, 46.2]	29.0 [26.6, 31.5]	1.672 [1.533, 1.816]	2.057 [1.957, 2.157]
AI scientists							
AI-Scientist-v2	✗	100.0	100.0	59.2 [55.1, 63.0]	43.8 [40.9, 46.8]	1.031 [0.947, 1.117]	1.448 [1.374, 1.522]
AutoDiscovery†	✗	100.0	100.0	27.5 [20.4, 35.8]	27.1 [20.4, 34.6]	1.474 [1.282, 1.663]	1.526 [1.287, 1.811]
Evaluation controls and MRs							
Empty control	-	100.0	100.0	16.7 [16.7, 16.7]	16.8 [16.7, 17.1]	1.935 [1.791, 2.087]	2.002 [1.803, 2.296]
Shuffled control	-	100.0	100.0	38.1 [34.9, 41.6]	39.9 [36.9, 42.8]	2.916 [2.741, 3.087]	3.146 [2.905, 3.414]
QMR	-	100.0	100.0	87.6 [84.9, 90.2]	82.9 [80.4, 85.3]	0.458 [0.406, 0.518]	0.538 [0.493, 0.586]
PMR	-	100.0	100.0	96.8 [95.8, 97.8]	97.5 [96.7, 98.2]	0.165 [0.118, 0.213]	0.133 [0.099, 0.172]

E.3 Agent Harnesses and Scientific Skills

Agent harnesses. We compare Codex and Claude Code using the same model backend and skill setting. Table 13 reports paired MCA differences on the same tasks. Claude Code generally achieves higher MCA, with the clearest difference observed for DeepSeek-V4-Flash without skills on easy tasks. The remaining confidence intervals include zero, so the observed differences do not establish a uniform advantage across configurations. These comparisons assess complete agent performance, including the empty-control fallback for invalid or missing outputs.

Table 13: Paired MCA differences between Claude Code and Codex. Positive values favor Claude Code. Brackets denote 95% task-bootstrap confidence intervals.

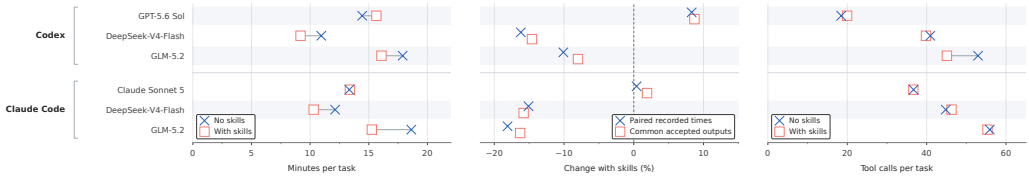
Model	Skills	Easy	Hard
DeepSeek-V4-Flash	✗	7.8 [2.2, 13.3]	2.9 [-0.9, 6.9]
DeepSeek-V4-Flash	✓	5.1 [-0.6, 10.7]	1.9 [-1.6, 5.2]
GLM-5.2	✗	4.8 [-1.0, 10.3]	0.1 [-2.7, 3.0]
GLM-5.2	✓	-0.3 [-5.2, 4.8]	0.8 [-2.2, 3.8]

The harnesses also differ in execution behavior. Claude Code makes more tool calls than Codex in all four shared-model comparisons (Figure 6). It takes longer with DeepSeek-V4-Flash under both skill settings. With GLM-5.2, Claude Code is slightly slower without skills but faster with skills. Thus, the observed MCA advantage is accompanied by more tool interactions, but not a consistent runtime advantage.

Scientific skills. Table 14 reports paired changes when scientific skills are made available. All MCA intervals include zero. MCNLL changes also generally have intervals spanning zero; Codex with DeepSeek-V4-Flash on hard tasks shows an increase whose interval excludes zero, indicating worse performance on this metric. These comparisons provide no consistent evidence that skill access improves the generated hypotheses. Skill access nevertheless changes execution time. Figure 6 shows mean tool calls and mean agent execution time, excluding environment setup and verification overhead. Comparisons use the same tasks with recorded execution times in both settings. DeepSeek-V4-Flash and GLM-5.2 run faster with skills under both harnesses, with reductions of 10.1-18.1%. Restricting the comparison further to tasks with accepted outputs in both settings retains reductions of 8.0-16.3%, indicating that early failures alone do not explain the pattern. GPT-5.6 Sol takes longer with skills, while Claude Sonnet 5 shows little runtime change.

Table 14: Paired performance changes with skills, computed as with skills minus without skills. MCA differences are in percentage points; positive MCA and negative MCNLL differences favor skills. Brackets denote 95% task-bootstrap confidence intervals.

Model	Easy / MCA	Hard / MCA	Easy / MCNLL	Hard / MCNLL
Codex				
GPT-5.6 Sol	0.3 [-5.0, 5.6]	1.5 [-1.8, 4.8]	0.001 [-0.122, 0.128]	-0.062 [-0.151, 0.025]
DeepSeek-V4-Flash	-1.2 [-6.0, 3.5]	-1.2 [-4.6, 2.4]	0.070 [-0.102, 0.240]	0.162 [0.006, 0.316]
GLM-5.2	1.6 [-3.5, 6.8]	-0.5 [-3.5, 2.5]	-0.035 [-0.199, 0.130]	0.000 [-0.226, 0.309]
Claude Code				
Claude Sonnet 5	-1.6 [-6.3, 3.3]	0.0 [-3.3, 3.2]	0.089 [-0.046, 0.222]	0.082 [-0.033, 0.199]
DeepSeek-V4-Flash	-3.9 [-8.6, 0.9]	-2.2 [-6.0, 1.6]	0.082 [-0.099, 0.269]	0.163 [-0.007, 0.330]
GLM-5.2	-3.5 [-8.7, 1.7]	0.2 [-2.5, 3.0]	0.162 [-0.016, 0.344]	-0.010 [-0.142, 0.119]



(a) Mean Execution Time / s (b) Execution Time Changes (c) Mean Tool Calls

Figure 6: Comparisons of execution time and tool calls with scientific skills access.

Tool-call counts do not change uniformly with runtime. For Claude Code with DeepSeek-V4-Flash, skill access shortens execution while slightly increasing the number of calls. For Claude Code with GLM-5.2, execution becomes faster with little change in call volume. These results distinguish

elapsed execution time from interaction count: fewer tool calls are not a necessary explanation for the observed runtime reductions.

Together, the performance and execution results show that scientific skill access can shorten analysis runs without consistently improving mechanistic hypothesis generation. The comparisons concern skill availability; they do not establish which skills were consulted or which aspects of the guidance caused the observed changes.

E.4 AI Scientists

Matched-task comparison. Tables 15 compare AI-Scientist-v2, AutoDiscovery, and Codex with GPT-5.6 Sol on the same 40 tasks used to evaluate AutoDiscovery, comprising 20 easy and 20 hard tasks. All three systems use GPT-5.6 Sol, holding both the model backend and task membership fixed. Codex is evaluated with and without scientific skills. AI-Scientist-v2's full-benchmark results remain its primary assessment; the matched subset supports direct comparison with AutoDiscovery. Each score includes all 20 tasks in its difficulty group, using the corresponding empty-control result for invalid or missing outputs. Brackets report 95% task-bootstrap confidence intervals. Controls and MRs are recomputed on the same task subset.

Table 15: Matched results in the AutoDiscovery subset. All generated hypotheses use GPT-5.6 Sol.

System / MRs	Skills	Valid Rate (%)↑		MCA (%)↑		MCNLL↓	
		Easy	Hard	Easy	Hard	Easy	Hard
Codex	✗	95.0	100.0	47.1 [35.0, 59.2]	35.8 [28.3, 44.2]	1.306 [1.043, 1.564]	1.802 [1.572, 2.029]
Codex	✓	100.0	95.0	57.9 [45.8, 69.6]	33.8 [27.5, 40.4]	1.099 [0.835, 1.385]	1.765 [1.511, 2.057]
AI-Scientist-v2	-	100.0	100.0	50.4 [38.3, 62.1]	35.0 [24.6, 45.4]	1.250 [1.001, 1.495]	1.534 [1.294, 1.794]
AutoDiscovery	-	100.0	100.0	27.5 [20.4, 35.8]	27.1 [20.4, 34.6]	1.474 [1.282, 1.663]	1.526 [1.287, 1.811]
Empty control	-	100.0	100.0	16.7 [16.7, 16.7]	17.9 [16.7, 20.4]	2.003 [1.629, 2.432]	1.769 [1.537, 2.006]
Shuffled control	-	100.0	100.0	37.9 [30.8, 44.6]	40.0 [32.5, 47.5]	2.861 [2.464, 3.266]	2.866 [2.374, 3.410]
QMR	-	100.0	100.0	87.9 [80.4, 94.6]	82.1 [74.6, 88.8]	0.480 [0.328, 0.677]	0.541 [0.434, 0.657]
PMR	-	100.0	100.0	97.1 [94.6, 99.2]	96.2 [93.3, 98.8]	0.160 [0.062, 0.276]	0.179 [0.079, 0.299]

AI-scientist performance. On the matched easy tasks, AI-Scientist-v2 has higher MCA and lower MCNLL than Codex without skills, but lower performance than Codex with skills on both metrics. On hard tasks, its MCA is close to both Codex configurations, while its MCNLL is lower. These point estimates show that its relative performance depends on task membership and metric; its lower scores on the full benchmark do not imply that it performs worse on every subset. AutoDiscovery achieves lower MCA than AI-Scientist-v2 on both difficulty groups. Its easy-task MCNLL is also higher, whereas their hard-task MCNLL values are nearly identical. Compared with Codex without skills, AutoDiscovery has lower MCA on both groups but lower MCNLL on hard tasks. The two AI-scientist workflows therefore differ in categorical agreement and probability assignment, despite using the same model and tasks. Both AI-scientist systems produce valid outputs for all 40 tasks, so their gaps from the qualitative and PMRs cannot be attributed to missing-output fallback. Both remain substantially below the MRs on both metrics. Specialized scientific-discovery workflows can change performance, but their use alone does not eliminate the difficulty of formulating mechanistic hypotheses from the available observations.

E.5 Evidence Summaries and Structured Records

Experimental design. We examine two forms of support for generating mechanistic hypotheses: evidence summaries that make conditional relationships in the public observations explicit, and

structured records that help agents retain and integrate findings during analysis. Both settings use the original public observations without access to hidden responses, ground truth outcomes, or MRs. Hypotheses are evaluated using the same interpretation and scoring procedure as in the main experiments.

Evidence summaries. Evidence summaries are computed exclusively from the public observations. For each predictor-response pair, observations are grouped by the remaining factor levels and phase conditions. Within each group, the summary computes response means and contrasts between adjacent predictor levels and between the first and last levels. Missing or nonnumeric response values are excluded from the means. A contrast is retained only when both sides contain at least eight rows and have defined response means. For each predictor-response pair, the report includes up to 24 contrasts with the largest absolute differences. Each entry records the fixed conditions, compared levels, response means, difference, and row counts. The report also gives the numbers of positive and negative contrasts across all eligible comparisons, together with overall row counts, condition-group counts, and phase frequencies. These summaries expose conditional associations that agents might otherwise overlook; they do not identify which associations have a mechanistic interpretation.

Structured records. In this setting, agents maintain an intermediate record of local mechanistic claims supported by the public observations. Each entry has a stable identifier and records the response channels, applicable factor levels and phase or history conditions, relevant groups, direction and approximate magnitude, supporting evidence, predictions under specified interventions, a falsifying observation, and uncertainty or alternative branches. Incompatible claims are retained as separate conditional branches. Agents are instructed to update the same record after each major analysis step and reread it before producing their final hypotheses. Each hypothesis must cite at least two entries and preserve their scope, branch conditions, predictions, and falsifiers. Conflicting findings must remain explicit conditional alternatives rather than being reduced to a single global direction. The record supports hypothesis construction; the submitted hypothesis remains the sole scientific account supplied to the hypothesis interpreter.

Results on the complete task subsets. Table 16 reports results on all 40 tasks for each configuration. Missing or invalid outputs receive the corresponding empty-control predictions, following Appendix D.5.

Table 16: Evidence-summary and structured-record results on the complete task subsets. Valid outputs are counted out of 40. Brackets contain 95% task-bootstrap confidence intervals.

Setting	Valid outputs	MCA (%)↑	MCNLL↓
Codex / GPT-5.6 Sol			
Baseline	37	39.2 [32.3, 46.5]	1.532 [1.344, 1.722]
Structured records	40	41.0 [34.4, 48.1]	1.509 [1.325, 1.696]
Evidence summaries	36	49.2 [41.2, 57.1]	1.244 [1.044, 1.469]
Claude Code / DeepSeek-V4-Flash			
Baseline	30	30.8 [25.0, 37.3]	2.032 [1.794, 2.262]
Structured records	32	32.9 [27.3, 39.2]	2.053 [1.770, 2.366]
Evidence summaries	39	40.8 [34.2, 47.5]	1.788 [1.466, 2.220]

Evidence summaries increase MCA by 10.0 percentage points in both configurations. The paired 95% confidence intervals are $[-1.0, 21.0]$ for Codex with GPT-5.6 Sol and $[3.1, 16.9]$ for Claude Code with DeepSeek-V4-Flash. MCNLL also decreases in both configurations. Structured records produce smaller MCA changes of 1.9 and 2.1 percentage points, respectively, with paired confidence intervals that include zero; their MCNLL changes are likewise inconclusive.

Results on tasks with valid outputs in all settings. To separate changes in output validity from differences among the submitted hypotheses, Table 17 restricts each configuration to tasks with valid outputs in all three settings. This leaves 33 tasks for Codex with GPT-5.6 Sol and 25 for Claude Code with DeepSeek-V4-Flash. Every setting within a configuration is evaluated on this same restricted task set.

Table 17: Results restricted to tasks with valid outputs in all three settings. Brackets contain 95% task-bootstrap confidence intervals.

Setting	MCA (%)↑	MCNLL↓
Codex / GPT-5.6 Sol: 33 tasks		
Baseline	39.1 [31.6, 47.0]	1.575 [1.365, 1.784]
Structured records	40.7 [33.1, 48.2]	1.530 [1.320, 1.753]
Evidence summaries	53.3 [44.7, 61.9]	1.132 [0.940, 1.338]
Claude Code / DeepSeek-V4-Flash: 25 tasks		
Baseline	38.7 [31.0, 47.3]	1.969 [1.654, 2.264]
Structured records	39.0 [31.7, 47.0]	1.939 [1.594, 2.316]
Evidence summaries	42.0 [33.3, 51.7]	1.589 [1.280, 1.884]

On these restricted task sets, evidence summaries improve Codex's MCA by 14.1 percentage points, with a paired 95% confidence interval of $[2.8, 25.8]$. The corresponding change for Claude Code is 3.3 points, with an interval of $[-5.0, 12.7]$. MCNLL decreases by 0.443 for Codex and 0.380 for Claude Code; the paired intervals for these changes are $[-0.671, -0.219]$ and $[-0.694, -0.055]$, respectively. Structured records yield small changes in both metrics, with paired intervals that include zero. These results are consistent with evidence identification being one source of difficulty: making conditional relationships explicit can support more accurate mechanistic hypotheses. Structured records provide a means to preserve findings, but do not supply relationships that the agent has failed to discover or ensure that recorded claims are integrated into a coherent explanation. Their limited gains therefore do not establish finding retention as the sole bottleneck.

E.6 Hypothesis Selection Diagnostic

Diagnostic definition. The primary evaluation uses the hypothesis with the highest agent-assigned weight. To examine whether this selection overlooks more accurate hypotheses among those submitted, we compare it with retrospective selection of the hypothesis achieving the highest MCA on each task. The retrospective choice maximizes probability across all twelve matched comparisons, selecting one hypothesis for the entire task. Ties in MCA are resolved by lower MCNLL. Both metrics are reported for this same selected hypothesis. This diagnostic uses ground truth directions after hypothesis generation and does not represent a selection procedure available to the agent. For each configuration, both selection rules are evaluated on the same tasks with valid generated outputs. Tasks without valid outputs are excluded from this diagnostic because they provide no hypotheses to compare. Consequently, the highest-weight scores reported here can differ from the complete-task results in Appendix E.2.

Results. Table 18 reports MCA under the two selection rules and their difference. Retrospective selection increases MCA across all configurations, showing that agent-assigned weights do not consistently identify the most accurate submitted hypothesis. Table 18 reports the corresponding MCNLL values. Retrospective selection also reduces mean MCNLL in every configuration, although maximizing MCA does not guarantee a reduction in MCNLL on each individual task. For Codex with GPT-5.6 Sol and scientific skills, retrospective selection raises MCA from 67.9% to 74.8% on easy tasks and from 52.0% to 58.9% on hard tasks. MCNLL decreases from 0.859 to 0.728 and from 1.298 to 1.090, respectively. On those same valid-output task sets, the PMR achieves MCA of 96.9% and 97.4%, with MCNLL of 0.161 and 0.137. Substantial gaps therefore remain in both metrics even after retrospective selection. These results indicate that hypothesis selection contributes to the observed errors, but better selection alone cannot recover the performance of the MRs from the hypotheses already generated. Under the fixed interpretation procedure, even the hypotheses selected retrospectively leave some response changes incorrectly predicted or assign insufficient probability to the ground-truth direction. Improving the content of the generated hypotheses therefore remains necessary alongside improving their ranking.

Table 18: Hypothesis selection diagnostic on tasks with valid outputs, reporting MCA and MCNLL on the same tasks. MCNLL under Best is evaluated for this same MCA-selected hypothesis. For both metrics, Δ is Best minus Highest, computed before rounding.

Model	Skills	MCA (%)↑						MCNLL↓					
		Easy			Hard			Easy			Hard		
		Highest	Best	Δ	Highest	Best	Δ	Highest	Best	Δ	Highest	Best	Δ
Codex													
GPT-5.6 Sol	✗	66.2	73.9	7.7	50.6	60.0	9.4	0.899	0.732	−0.166	1.360	1.140	−0.220
	✓	67.9	74.8	6.9	52.0	58.9	6.9	0.859	0.728	−0.131	1.298	1.090	−0.208
DeepSeek-V4-Flash	✗	50.4	61.7	11.4	35.6	48.4	12.9	1.624	1.229	−0.395	2.126	1.509	−0.617
	✓	47.6	60.5	12.9	33.4	46.7	13.3	1.765	1.200	−0.565	2.247	1.571	−0.675
GLM-5.2	✗	49.3	62.7	13.4	33.9	48.5	14.6	1.573	1.075	−0.498	2.124	1.437	−0.687
	✓	49.6	62.6	12.9	32.7	47.0	14.2	1.497	1.075	−0.422	2.020	1.458	−0.561
Claude Code													
Claude Sonnet 5	✗	50.1	62.2	12.1	31.0	41.6	10.6	1.422	1.050	−0.372	1.898	1.577	−0.321
	✓	46.9	59.4	12.5	31.1	43.1	12.0	1.523	1.143	−0.380	1.981	1.531	−0.451
DeepSeek-V4-Flash	✗	51.8	62.7	10.9	36.0	50.7	14.7	1.587	1.169	−0.417	2.047	1.448	−0.599
	✓	46.8	62.1	15.3	33.2	47.9	14.7	1.712	1.169	−0.543	2.204	1.525	−0.679
GLM-5.2	✗	54.1	63.8	9.7	33.5	45.8	12.3	1.380	1.036	−0.344	2.075	1.555	−0.520
	✓	48.9	63.1	14.1	32.0	46.6	14.7	1.600	1.061	−0.539	2.145	1.565	−0.579
AI scientists													
AI-Scientist-v2	-	59.2	73.5	14.3	43.8	59.5	15.6	1.031	0.732	−0.299	1.448	1.091	−0.357
AutoDiscovery	-	27.5	35.4	7.9	27.1	35.0	7.9	1.474	1.370	−0.105	1.526	1.198	−0.327

F Case Studies

The following examples show public data, excerpts from the highest-weight hypotheses, and their evaluation. Response channels are generated from paper-derived mechanisms on independently collected datasets. In the data excerpts, group identifiers are shortened locally to G1, G2, and so forth; time denotes `observation_time_value`, and history denotes `factor_gamma_level`. Response values are rounded to two decimals; record identifiers, covariates, and other columns are omitted. Italicized passages reproduce agent text, with ellipses marking omissions within a passage.

F.1 Successful Mechanistic Hypothesis

In Figure 7, the hypothesis connects a signed interaction with phase-dependent activation and reserve components. In the selected comparisons, its phase-dependent interaction supports the increase in C in q03, while its restriction of history effects to the reserve component supports no change in C in q08. The declining reserve contribution under low Alpha supports the decrease in D in q09. These predictions follow relationships stated in the hypothesis, linking its mechanistic content to the successful evaluation.

A successful hypothesis on a hard task

Task: `task_897d14e1d81d`. Codex / GPT-5.6 Sol, without skills; hard task. Highest-weight hypothesis: 12/12 correct comparisons, MCA 100.0%, MCNLL 0.178.

Scientific sources: The mechanism is derived from *Transgenerational plasticity of inducible defences: Combined effects of grand-parental, parental and current environments (Tariel et al., 2020)*. The independent Zenodo dataset contains poultry production and egg-quality measurements (DOI `10.5281/zenodo.19841366`). Treatment, feed consumption, weight change, and day supply Alpha, Beta, Gamma history, and phase.

Public data (excerpt): First three of 6,857 rows in `data_001.xlsx`; selected columns.

Group	Time	Alpha	Beta	Gamma history	Phase	C	D	E
G1	1	low	mid	mid	early	44.53	67.58	62.81
G1	2	low	mid	mid	early	48.01	69.65	64.95
G1	23	low	mid	low	middle	36.91	51.39	61.04

Generated hypothesis (excerpts):

"Let P be a slowly accumulating activation arm and Q an opposing reserve arm. Channel A reads P: Alpha raises A, and that Alpha effect grows from early through late and remains elevated, though flatter, in recovery. Channel B reads Q: Alpha lowers B strongly in early observations, the Alpha contrast weakens later, and low-Alpha B also decays with time. ... A separate latent interaction state S drives Channel C. ... Thus, increasing Beta suppresses C when Alpha is low but enhances C when Alpha is high; the divergence becomes largest late or in recovery. Channel C is the upper-censored measurement clip(S,100). Channel D is a downstream sum of the same uncensored S and the reserve arm: D = clip(S + B - 50 + independent noise,100). ... Factor Gamma history is a weak compensatory memory acting mainly on Q, not a current-state driver: high lagged Gamma raises B and D at low Alpha but modestly lowers them at high Alpha, with its effect strongest early and fading later."

Withheld comparisons: Directions describe the first condition relative to the second.

ID	Comparison	Ground Truth	Interpreter
q03	Channel C at middle versus early phase, with high Alpha, high Beta, and mid Gamma history.	Increase	Increase
q08	Channel C at mid versus high Gamma history, with high Alpha, low Beta, and early phase.	No change	No change
q09	Channel D at recovery versus late phase, with low Alpha, low Beta, and high Gamma history.	Decrease	Decrease

Figure 7: A successful hypothesis on a hard task, with public data and withheld comparisons.

F.2 Missing a Conditional Relationship

In Figure 8, the hypothesis treats an approximately additive marginal pattern as phase-invariant. With function names shortened, its Channel B equation is $B = b_0 + f_{\mathrm{Beta}}(\mathrm{Beta}) + f_{\mathrm{Gamma}}(\mathrm{Gamma}) + \epsilon_B$, which contains no phase term. The public conditional contrasts show a different pattern: at high Alpha and low Beta, the mean recovery-minus-early change in B is -26.665 under low Gamma history but $+23.110$ under high Gamma history. At mid Alpha and low Beta, the corresponding changes are -26.659 and $+23.949$. Thus, the phase effect reverses with history while Alpha and Beta are fixed. Although the hypothesis identifies phase-dependent residuals as a potential falsifier, this dependence is absent from its submitted equation. For q04, q06, and q08, the interpreter follows that equation and predicts no change. The errors therefore reflect the omitted conditional relationship rather than a failure to apply the stated hypothesis.

A hypothesis that misses a conditional effect

Task: `task_584e0b187c00`. Codex / DeepSeek-V4-Flash, without skills; hard task. Highest-weight hypothesis: 3/12 correct comparisons, MCA 25.0%, MCNLL 2.701.

Scientific sources: The mechanism is derived from *Self-regulation strategy, feedback timing and hemodynamic properties modulate learning in a simulated fMRI neurofeedback environment (Oblak et al., 2017)*. The independent OSF longitudinal dataset (DOI `10.17605/osf.io/8efud`) supplies session attendance, social support, baseline symptom score, and visit month as Alpha, Beta, Gamma history, and phase.

Public data (excerpt): First three of 5,119 rows in `data_001.csv`; selected columns. The full table contains 1,042 groups.

Group	Time	Alpha	Beta	Gamma history	Phase	B	D
G1	0	low	low	mid	early	41.94	31.37
G1	1	low	mid	mid	middle	46.70	40.47
G1	6	low	mid	mid	middle	51.76	52.10

Generated hypothesis (excerpts):

"The first stream, governing channels B and C, implements an instantaneous additive combination of two factor inputs: the current beta (b) and the block-level gamma (g). ... The beta and gamma effects are additive with negligible interaction. ... This hypothesis is falsified if ... the B and C residuals after the additive beta+gamma model show systematic patterns with phase (indicating the instantaneous channels also integrate)"

Withheld comparisons: Directions describe the first condition relative to the second.

ID	Comparison	Ground Truth	Interpreter
q04	Channel B at middle versus recovery phase, with high Alpha, mid Beta, and low Gamma history.	Increase	No change
q06	Channel B at late versus middle phase, with high Alpha, mid Beta, and low Gamma history.	Decrease	No change
q08	Channel B at late versus middle phase, with mid Alpha, mid Beta, and low Gamma history.	Decrease	No change

Figure 8: A hypothesis that misses a conditional phase effect, with public data and withheld comparisons.

F.3 Effects of Evidence Summaries and Structured Records

In Figure 9, the baseline mentions an Alpha effect on the total CDE pool, but its final CDE equations omit Alpha and Gamma history, leaving their response paths incompletely specified. The structured-record hypothesis incorporates the negative history effect on D and its contribution to E, correcting q01, while leaving Alpha to another hypothesis and missing q07 and q12. With evidence summaries, the response equations connect Alpha to D and E through the shared pool and condition the Gamma effect on Beta. Although the text retains the language of independent axes, these explicit paths support all three directions shown and all twelve comparisons overall. The example distinguishes retaining a local finding from incorporating the relevant relationships into the selected standalone explanation.

How evidence support changes a hypothesis

Task: `task_fcc33a74fab2`. Claude Code / DeepSeek-V4-Flash, without skills; easy task. Settings follow Appendix E.5.

Scientific sources: The mechanism is derived from *A positive feedback loop mediates crosstalk between calcium, cyclic nucleotide and lipid signalling in calcium-induced Toxoplasma gondii egress (Nofal et al., 2022)*. The independent dataset is a rural child mental-health survey from ScienceDB (DOI `10.57760/sciencedb.j00001.00464`), with 1,224 public rows and 127 covariates.

Public data (excerpt): First three rows of `data_001.xlsx`; selected columns. The same observations are available in all three settings.

Group	Time	Alpha	Beta	Gamma history	Phase	C	D	E
G1	4	high	mid	low	late	24.73	57.17	56.15
G2	3	high	mid	mid	late	24.43	50.08	55.09
G3	5	low	low	low	recovery	36.16	37.30	42.33

Baseline hypothesis (excerpts): 5/12 correct; MCA 41.7%; MCNLL 2.506.

"The factor_alpha_level has a secondary effect on C through an interaction ... This means alpha affects the total CDE pool size slightly, not the allocation ratio."

Structured-record hypothesis (excerpts): 7/12 correct; MCA 58.3%; MCNLL 1.610.

"The second subsystem (Channels A and B) is driven by Factor Alpha and covered in Hypothesis 2. ... The Gamma history state provides a mild negative correction: each unit of Gamma value reduces D by 0.45 units."

Evidence-summary hypothesis (excerpts): 12/12 correct; MCA 100.0%; MCNLL 0.135.

*"The total resource pool P = P_0 + k_alpha*alpha + k_beta*beta + epsilon ... is not conserved; it grows with both factors. The allocation variable L = f(beta, gamma, phase) determines the fraction of pool P that flows to the d/e side versus the c side."*

Withheld comparisons and interpreter predictions: Each comparison describes the first condition relative to the second. Bold predictions agree with the ground-truth direction.

Comparison	Ground Truth	Baseline	Structured records	Evidence summaries
q01: E at low versus high history; mid Alpha, low Beta, late phase.	Increase	No change	**Increase**	**Increase**
q07: D at high versus low Alpha; high Beta, low history, late phase.	Increase	No change	No change	**Increase**
q12: D at low versus high Alpha; high Beta, low history, middle phase.	Decrease	No change	No change	**Decrease**

Figure 9: Hypotheses for the same task under baseline, structured-record, and evidence-summary settings. Scores cover all twelve comparisons; three are shown.