SCIGAME: A CONTROLLED BENCHMARK FOR EVIDENCE-DRIVEN SCIENTIFIC LAW DISCOVERY IN LANGUAGE MODEL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model agents are increasingly applied to scientific discovery, yet existing benchmarks conflate law inference with programming, assess static fits over unseen numerical prediction, and fail to separate empirical induction from pre-trained priors. We introduce SCIGAME, a controlled benchmark that distills literature into structured scientific cards to instantiate paired Original and dimensionally consistent Counterfactual environments under an identical task interface. Spanning canonical Historical laws alongside a live Temporal collection dynamically curated from post-cutoff peer-reviewed papers, SCIGAME systematically benchmarks agents across both candidate-guided Identification and open-ended Discovery settings under prior-only reasoning, fixed observation batches, and adaptive sequential experimentation protocols. Evaluating five frontier models reveals three core findings: (1) rooted priors become liabilities once governing relations are altered; (2) adaptive experimentation benefits the strongest models but degrades weaker ones under constrained budgets; and (3) high explanation quality can mask quantitative failures, with exact symbolic formulation remaining the primary bottleneck even when agents reliably interpret empirical evidence. These results demonstrate that while current agents can actively investigate systems, reliably inducing the exact physical laws that govern unfamiliar worlds remains an open challenge.

1 INTRODUCTION

Scientific discovery is a sequential process in which researchers formulate hypotheses, design experiments, interpret observations, and derive conclusions that remain valid within clearly specified conditions (). Large language models (LLMs) are increasingly capable of participating in parts of this process. Recent systems can analyze scientific descriptions (M. Bran et al., 2024; Laurent et al., 2024), construct executable mathematical expressions (Shojaee et al., 2025; Lu et al., 2024), interact with simulated environments (Wang et al., 2022), and propose experiments to test competing hypotheses (Boiko et al., 2023; Huang et al., 2025; Gottweis et al., 2026). These capabilities suggest that LLM agents may assist with scientific discovery beyond conventional question answering or data fitting.

Scientific law discovery, however, requires more than reproducing a relation from observed data (Schmidt & Lipson, 2009; Brunton et al., 2016; Udrescu & Tegmark, 2020). An agent must infer how controllable inputs determine observables, generalize the inferred relation to unqueried inputs (Kronberger et al., 2022), distinguish evidence from prior knowledge (Kronberger et al., 2022; Ríos-García et al., 2026), revise its hypotheses when observations contradict familiar laws (Wilie et al., 2024; Jhaveri et al., 2026), and explain both the discovered relation and its applicability conditions. A reliable benchmark must therefore evaluate the complete investigation process rather than only the final equation or the quality of a generated explanation (Ríos-García et al., 2026).

Recent benchmarks have made important progress toward this goal. Interactive environments such as Discovery-World (Jansen et al., 2024), FalsifyBench (Bertolazzi et al., 2026), and Auto-Discovery-Bench (Chen et al., 2025a) evaluate experiment selection, hypothesis revision, and rule discovery. Equation-discovery benchmarks such as LLM-SRBench Shojaee et al. (2025) focus on recovering mathematical relations from data, while BoxingGym (Gandhi et al., 2025) and ActiveSciBench-Chem (Kabra et al., 2026) study the connection between model construction and active experimental design. Other benchmarks, including Gravity-Bench (Koblischke et al., 2025), PhysGym Chen et al. (2025b), NewtonBench (Zheng et al., 2026), and DiscoverPhysics (Wiemann et al., 2026), examine scientific reasoning under supplied context or altered physical laws. Paper-grounded benchmarks such as DiscoveryBench (Majumder et al., 2025), ScienceAgentBench (Chen et al., 2025c), PaperBench (Starace et al., 2025), and SciLaws-Bench (Huang et al., 2026a) extend evaluation to scientific sources, datasets, and reported results.

These benchmarks nevertheless leave several methodological gaps. First, they often vary the task, observation protocol, or evaluation criterion at the same time, which makes it difficult to determine whether a performance difference comes

from empirical observations, supplied candidate relations, or active experiment selection. In particular, few benchmarks compare prior-only reasoning, fixed observations, and adaptive experimentation on the same task interface while also pairing an original relation with a controlled counterfactual relation (Chen et al., 2025b; Zheng et al., 2026; Wiemann et al., 2026). Second, existing evaluations frequently emphasize symbolic recovery, fitting quality, or task completion. Such measures do not always test numerical generalization to unseen inputs, and a scientifically plausible explanation may receive a high score even when the submitted predictor is quantitatively incorrect (Ríos-García et al., 2026). Third, most scientific benchmark collections are static. They provide limited support for evaluating models on newly published scientific relations that were unavailable at the model's knowledge cutoff (White et al., 2025; He et al., 2026). Finally, paper-based research tasks often combine law discovery with programming, data analysis, or research replication, making it difficult to isolate the ability to infer and test a scientific relation (Huang et al., 2024; Kon et al., 2026).

We introduce SCIGAME, a controlled benchmark for evaluating whehter LLM can discovery scientific law. The benchmark converts source relations into structured scientific cards that record the relation, parameters, variable definitions, units, valid input domains, and applicability conditions. Each card gives rise to a pair of simulator-backed environments. The *Original* environment implements the source relation, whereas the *Counterfactual* environment applies a controlled, dimensionally consistent mutation to that relation. The paired environments share the same task interface, input ranges, observables, and experimental constraints. They therefore make it possible to test whether an agent follows empirical evidence when the active relation differs from a familiar scientific law (Zheng et al., 2026; Wiemann et al., 2026).

Each environment supports two complementary task settings. In *Identification*, the agent receives candidate mathematical expressions and must determine which relation governs the environment. In *Discovery*, candidate expressions are withheld and the agent must construct a predictor from the task description and experimental evidence. Both settings require an executable expression for every observable, a scientific explanation, and explicit applicability conditions.

To separate the effects of prior knowledge, static observations, and experimental autonomy, every task can be evaluated under three protocols. Under *Prior-only*, the agent receives no empirical observations. Under *Fixed*, the framework provides a batch of observations generated by a prespecified experimental design. Under *Adaptive*, the agent begins without observations, proposes experiments, receives observations or constraint feedback through an `Oracle`, updates its hypotheses, and decides when to submit its final predictor. All three protocols use the same task interface and final-output contract, while differing in how observations are obtained. This design supports a controlled comparison between relying on scientific priors, using predetermined evidence, and autonomously selecting experiments.

SCIGAME also includes a live, knowledge-cutoff-aware benchmark track. Its Historical collection contains stable scientific relations that provide interpretable reference tasks. Its Temporal collection is constructed through a rerunnable curation pipeline that verifies publication and preprint dates, extracts quantitative relations from eligible papers, and subjects the resulting scientific cards to independent review. The temporal window can be shifted as model knowledge cutoffs change, enabling evaluation on scientific relations that were unavailable during pretraining and reducing the risk that performance is explained by memorization (Rein et al., 2023; White et al., 2025).

We evaluate the agent from three complementary perspectives. A prediction score measures numerical accuracy on hidden evaluation inputs. A hierarchical episode success rate measures end-to-end completion under both numerical-tolerance and exact-expression standards. A rubric-based scientific explanation score evaluates whether the agent correctly describes the relation, derives supported consequences, and limits its claims to the available evidence and applicability conditions (Zheng et al., 2023; Tan et al., 2025). The recorded interaction history further enables analysis of experimental choices and failure modes.

Our contributions are as follows:

- We construct a controlled scientific-law discovery benchmark with source-grounded scientific cards, paired Original and Counterfactual environments, and both Identification and Discovery tasks.
- We introduce a live Temporal collection with a knowledge-cutoff-aware curation pipeline that can be updated as new scientific relations become available.
- We provide a unified evaluation framework that compares Prior-only, Fixed, and Adaptive protocols using hidden-input prediction accuracy, hierarchical episode success, and scientific explanation quality.
- Through experiments across multiple language models, we find that familiar scientific priors help on Original relations but can hinder performance on Counterfactual relations; adaptive experimentation benefits stronger models but can harm weaker models under limited interaction budgets; incorrect functional-relation discovery remains the dominant failure mode; and high-quality scientific explanations can coexist with quantitatively incorrect predictors. The Anonymous code and dataset is available at this anonymous repository.

Table 1: **Comparison of scientific discovery benchmarks and our work.** ✓ indicates supported and — unsupported. Paper grounding covers paper-derived tasks, Counterfactual includes altered relation variants, and Temporal selection denotes date-based source curation.

Benchmark	Active experiments	Numerical prediction	Paper grounding	Counterfactual	Temporal selection
DiscoveryWorld (Jansen et al., 2024)	✓	✓	—	—	—
FalsifyBench (Bertolazzi et al., 2026)	✓	—	—	—	—
Auto-Discovery-Bench (Chen et al., 2025a)	✓	✓	—	—	—
LLM-SRBench (Shojaee et al., 2025)	—	✓	—	✓	—
BoxingGym (Gandhi et al., 2025)	✓	✓	✓	—	—
ActiveSciBench-Chem (Kabra et al., 2026)	✓	✓	—	✓	—
Gravity-Bench (Koblischke et al., 2025)	✓	—	✓	✓	—
PhysGym (Chen et al., 2025b)	✓	—	—	—	—
NewtonBench (Zheng et al., 2026)	✓	✓	—	✓	—
DiscoverPhysics (Wiemann et al., 2026)	✓	✓	—	✓	—
DiscoveryBench (Majumder et al., 2025)	—	—	✓	—	—
ScienceAgentBench (Chen et al., 2025c)	—	✓	✓	—	—
PaperBench (Starace et al., 2025)	✓	✓	✓	—	✓
SciLaws-Bench (Huang et al., 2026a)	✓	✓	✓	✓	—
SciGame (**Ours**)	✓	✓	✓	✓	✓

2 Relevant work

Scientific discovery benchmarks. Scientific discovery benchmarks evaluate how agents infer relations from observations and design experiments to test their hypotheses. Interactive benchmarks such as DiscoveryWorld (Jansen et al., 2024), FalsifyBench (Bertolazzi et al., 2026), and Auto-Discovery-Bench (Chen et al., 2025a) study hypothesis revision and experiment selection. LLM-SRBench (Shojaee et al., 2025), BoxingGym (Gandhi et al., 2025), and ActiveSciBench-Chem (Kabra et al., 2026) evaluate equation discovery and active experimental design. Other benchmarks vary the scientific context or underlying laws, including Gravity-Bench (Koblischke et al., 2025), PhysGym (Chen et al., 2025b), NewtonBench (Zheng et al., 2026), and DiscoverPhysics (Wiemann et al., 2026). However, they do not compare the same prediction task with and without supplied candidate expressions, leaving unclear how these expressions affect predictive accuracy when observations support or contradict a reference relation.

Paper-based scientific research. Paper-grounded benchmarks extend scientific-discovery evaluation to questions, datasets, and reported results. DiscoveryBench (Majumder et al., 2025) and ScienceAgentBench (Chen et al., 2025c) assess hypothesis discovery and scientific programming, while PaperBench (Starace et al., 2025) evaluates computational research replication. Their success criteria therefore reflect broader research skills than relation inference alone. SciLaws-Bench (Huang et al., 2026a) directly targets paper-derived law discovery, but its Real and Parallel settings use different observation and evaluation protocols, preventing a matched comparison of predictive accuracy.

Table 1 summarizes feature coverage across the reviewed benchmarks. SciGame is the only one that jointly considers all these features.

3 Methodology

3.1 Overview

Figure 1 summarizeds our evaluation pipeline and adaptive protocol for scientific investigation. We distill information from scientific sources into structured scientific cards, each recording a relation, parameter values, variable definitions, units, a valid input domain, and applicability conditions. Each card is used to instantiate an Original task environment and a paired Counterfactual environment. The two environmnets share the same task interface and experimental constraints, while the Counterfactual uses a controlled structural mutation of the underlying relation. An episode is one run of a `Player` on a single task environment. During an episode, the `Player` proposes experiments to an `Oracle`, receives observations or feedback, and updates a set of hypotheses and associated confidence values until declaring completion. We evaluate the final predictor and its scientific explanation and analyze the recorded interaction trace, thereby assessing both task performance and investigation behavior. Section 3.2 describes source-relation curation and paired-environment constructionm of paired environments; Section 3.3 details the experimental protocol; and Section 3.4 defines the proposed evaluation metrics.

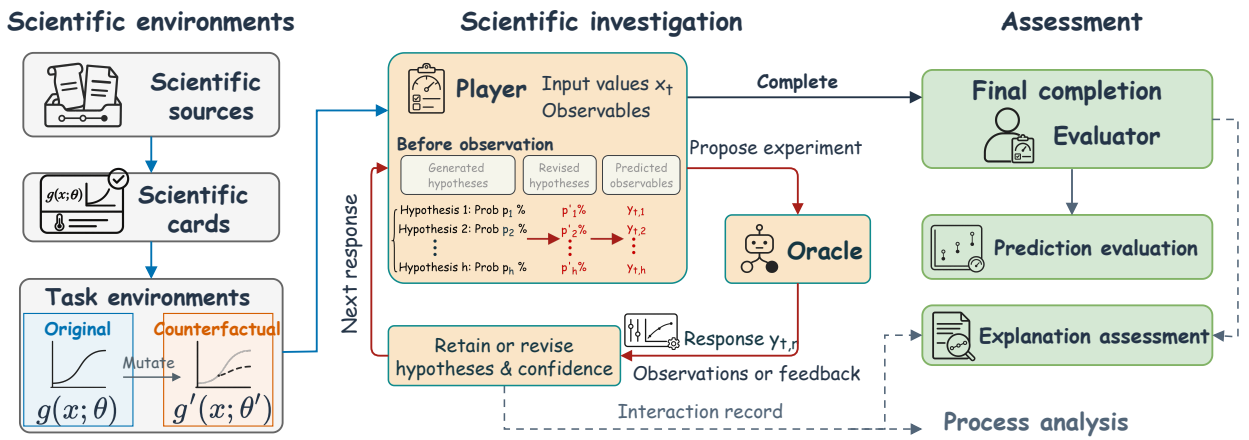


Figure 1: **Overview of the study pipeline.** Scientific cards record the relations, variables, and applicability conditions used to construct the simulated environments. The central panel illustrates the **Adaptive** protocol, where the `Player` dynamically designs experiments and autonomously decides when to complete. Final predictions, scientific explanations, and recorded interaction trajectories are evaluated post-episode.

Problem formulation. Each task specifies controllable inputs and observables, their respective units, and the conditions under which experiments can be performed. A deterministic simulator defines the response $g_r(x)$ of observable r to an input vector x within the task's valid input domain. The simulator's underlying relation and fixed parameters remain unchanged throughout an episode, which is defined as a single run of the `Player` on a specific task. The `Player` constructs a predictor $\hat{g}$, represented by executable expressions $\hat{g}_r(x)$ for all task observables, using the task description, any provided candidate expressions, and observations obtained according to the experimental protocol. Finally, its scientific explanation detail the inferred relation and the conditions under which the model applies.

3.2 Data Curation

Scientific data sources. We compile two distinct collections of source relations. The **Historical** collection comprises relations drawn from classical scientific history, serving as interpretable references for examining whether agents can use established scientific relations to make experimental predictions and justify their conclusions with evidence. The **Temporal** collection extracts relations from scientific publications within a configurable time window, which is setup according to the evaluated `Player`'s reported knowledge cutoff . This approach reduces the risk of data contamination when evaluating contempory `Player`s.

Scientific data construction. For each task family, we create a scientific card recording the selected relation, parameter values, variable definitions, units, input domain, and applicability conditions. Historical relations are formulated under fixed modeling assumptions with explicitly defined controllable inputs and observables. For the Temporal collection, a live curation pipeline (Figure 2) retains only papers whose verified journal publication and initial preprint dates both fall strictly after the evaluated `Player`'s reported knowledge cutoff. An LLM extracts quantitative relations from these papers, and an independent LLM reviewer checks each card for accuracy and completeness. Cards with missing or unsupported information are revised or excluded. The pipeline can be rerun with updated date windows to refresh the benchmark over time. All relevant prompts are provided in Appendix A.

Counterfactual construction. For each task family in both collections, we pair an Original environment, which uses the source relation, with a Counterfactual environment obtained by systematically mutating that relation. The two environments share the same controllable inputs, permitted input ranges, observables, and experimental constraints; they differ only in the mathematical relation used to generate observations. This controlled pairing tests whether the `Player` uses empirical evidence to adapt when observations conflict with established scientific priors, rather than relying solely on memorized knowledge. Following the counterfactual construction framework introduced by NewtonBench (Zheng et al., 2026), we apply structural mutations, including changes to coefficients and exponents, nonlinear transformations, and cross-input interactions. To preserve physical validity, we maintain dimensional consistency during nonlinear transformations and update any mutated variable consistently across all related equations. Detailed mutation protocols for the Historical and Temporal families are provided in Appendix C. Finally, we cross-evaluate the exact ground-truth equations of each environment on observations from its counterpart and verify that they fail the prediction evaluation, confirming that the paired environments are empirically distinguishable.

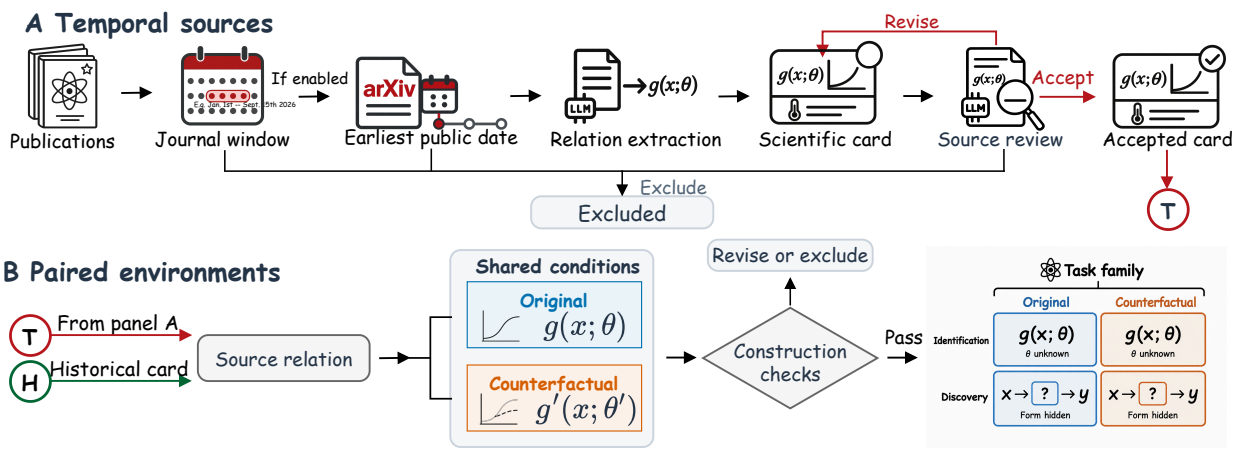


Figure 2: **Data curation pipeline**. Part A describes the workflow on how to formulate accepted scientific cards from temporal sources. Part B describes how both historical and temporal corpus be curated as task family collections.

3.3 Experimental framework and protocol

For each Original and Counterfactual environment, we define two task settings:

Identification. The `Player` agent receives a formula that combines the Original and Counterfactual expressions using one unknown parameter s, such that $G_r(x; s) = (1 - s) \cdot g_r(x) + s \cdot g'_r(x)$.when $s = 0$, the Original expression $g_r(x)$ is selected; conversely, $s = 1$ select the Counterfactual one $g'_r(x)$. The task specifies whether intermediate parameter values are also allowed, in which case the formula linearly interpolates between the two predictions. While the simulator fixes this parameter strictly at zero or one, the active value remains hidden from the `Player`.

Discovery. The `Player` receives neither candidate expressions nor predefined parametric models. Instead, it must autonomously construct a mathematical relation mapping experimental inputs to observables based purely on the initial background description, experimental constraints, and available observations.

Combining the Original and Counterfactual environments with these two settings yields four distinct tasks per scientific family. Both settings require executable expressions for the observables and a scientific explanation. Furthermore, both settings share the same experimental interface and numerical evaluation criteria within a given environment.

An episode is a single run of the `Player` on one task under one protocol. Each episode begins with an empty interaction history and is isolated from all other episodes. As shown in Figure 3, We evaluate three protocols including **Prior-only**, **Fixed**, and **Adaptive**, whose differences are detailed in below.

Prior-only baseline. Under the Prior-only protocol, the `Player` receives no empirical observations and does not interact with the `Oracle`. It relies only on the task description, protocol instructions, and prior scientific knowledge before producing its final completion.

Fixed experimental design. Under the Fixed protocol, the framework provides observations generated by a prespecified experimental design with a fixed budget. The `Player` cannot select, modify, or repeat experiments. After receiving these observations, it produces a final completion in the same format as the Prior-only protocol.

Adaptive experimental design. Under the Adaptive protocol, the `Player` begins with zero observations, then may iteratively propose experiments to the `Oracle`. Each proposal contain at least two hypotheses with nonnegative confidence scores that sum to one. For every proposed input and requested observable, the `Player` must provide a numerical prediction under each hypothesis.

At each LLM call, the `Player` receives the task description, protocol instructions, available observations, and the current interaction history. It returns a structured response containing its hypotheses, a confidence score for each hypothesis, and either an experimental proposal or a final completion. Every final completion must include a predictor, a scientific rationale, and the conditions under which the predictor applies.

The `Oracle` validates each proposal against the task's experimental constraints. For an accepted proposal, it returns simulated observations; for a rejected proposal, it returns constraint-related feedback without revealing observational data. The framework appends the proposal and the `Oracle`'s response to the interaction history before the next LLM call. The `Player` then updates its hypotheses and either proposes another experiment or submits a final completion. The Adaptive loop terminates when the confidence of the most likely hypothesis exceeds a predefined threshold τ.

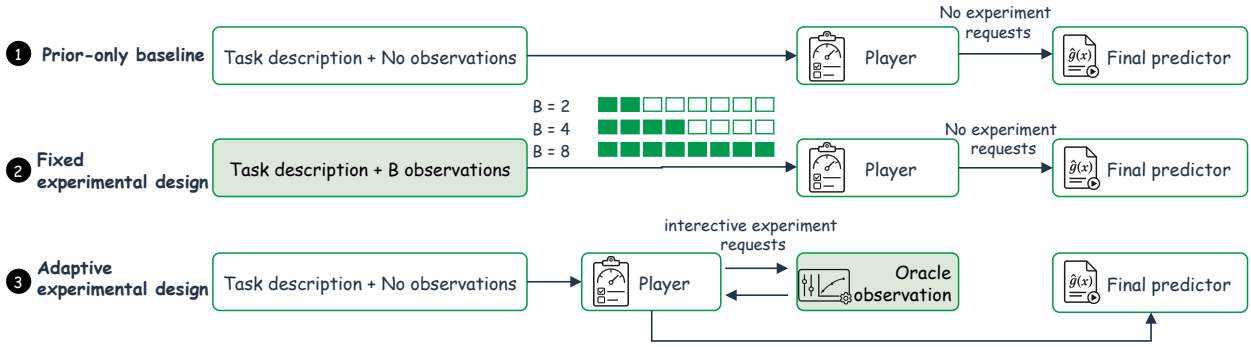


Figure 3: **Comparison of observation protocols.** All three methods share the initial task description and assigned model information. **Fixed** provides a batch of observations from B pre-specified experiments upfront. **Adaptive** allows sequential experiment selection without a pre-set budget constraint.

3.4 Evaluation metrics

We define three metrics to evaluate the `Player`'s ability to discover scientific laws. The prediction score measures the numerical accuracy and generalization of the submitted predictor on unseen inputs. The episode success rate measures whether the `Player` completes the task successfully under numerical-tolerance and exact-expression criteria. The scientific explanation score evaluates whether the `Player` correctly explains the discovered relation, its consequences, and its applicability conditions. Unless otherwise stated, we report the number of valid or determinable episodes contributing to each metric and account for invalid or missing results separately.

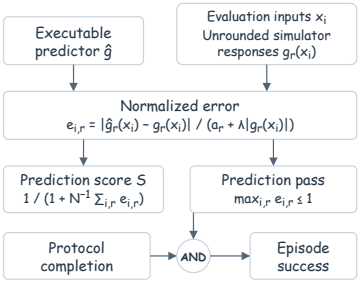


Figure 4: **Evaluation pipeline.**

Prediction score. To test generalization, the `Evaluator` compares the submitted mathematical predictions against the simulator's ground-truth responses using a hidden set of evaluation inputs unseen by the `Player`. For all subsequent analyses, we explicitly report the number of valid episodes contributing to each metric and account for invalid or missing results separately. As shown in Figure 4, for an episode we normalize the absolute prediction errors by the declared tolerances and compute:

$$e_{i,r} = \frac{|\hat{g}_r(x_i) - g_r(x_i)|}{a_r + \lambda|g_r(x_i)|}, \text{and } e = \frac{1}{N}\sum_{i,r} e_{i,r}, \quad S = \frac{1}{1+\bar{e}}, \tag{1}$$

where x_i is an evaluation input, a_r and λ are the absolute and relative tolerances, and N is the number of evaluated input and observable pairs. Higher scores indicate more accurate predictions, with $S = 1$ for zero error. This continuous score also distinguishes completions that fail the tolerance requirement but differ in predictive accuracy.

Episode success rate. We here establish a **hierarchical** episode success standard. **The weak standard** is illustrated in Figure 4, where an episode succeeds only when the prediction satisfies $\max_{i,r} e_{i,r} \le 1$ on preset inputs x and the completion satisfies the protocol. **The strict standard** further requires the final predictor $\hat{g}$ to be identical to the ground-truth expression. The episode success rate is $\text{ESR} = \frac{N_\text{s}}{N_\text{all}}$, where N_s counts successful episodes and N_all counts episodes with a determined success or failure outcome.

Scientific explanation score. After the `Player` runs are complete, we conduct a separate post hoc evaluation using an LLM agent as a judge. The judge assesses the `Player`'s scientific explanations from the task description and recorded interactions. We additionally provide the simulator's reference relation and applicability conditions as the basis for assessing scientific correctness.

These metrics provide complementary evaluations of scientific-law discovery. The prediction score measures quantitative generalization, the episode success rate measures end-to-end task completion, and the scientific explanation score measures the correctness and evidential scope of the `Player`'s scientific reasoning. We report all three metrics and the corresponding counts of valid, invalid, missing, and undetermined episodes. More detais are provided in Appendix A.2.

Table 2: **Overall benchmark performance on both collections.** Episode success rate (ESR, %), average prediction score (PS), and scientific explanation score (SES). Each cell pools three repetitions per family across all 20 Historical and 24 Temporal families, and Temporal families were published between 30 April and 15 September 2026. ESR is the **strict standard**, which requires the final predictor to be identical to the ground-truth expression. PS is the mean held-out prediction score, and SES is the mean judge score.

		Historical												**Temporal**									
		Identification						Discovery						Identification						Discovery			
		Original			Counterfactual			Original			Counterfactual			Original			Counterfactual			Original			Counterfactual		
Model	Method	ESR	PS	SES	ESR	PS	SES	ESR	PS	SES	ESR	PS	SES	ESR	PS	SES	ESR	PS	SES	ESR	PS	SES	ESR	PS	SES
`GPT-5.5`	Prior-only	73.3	0.780	0.884	31.7	0.404	0.519	35.0	0.655	0.939	0.0	0.123	0.289	8.3	0.244	0.517	19.4	0.389	0.768	0.0	0.209	0.587	0.0	0.081	0.530
	Fixed	95.0	1.000	0.997	95.0	1.000	0.983	63.3	0.998	0.943	0.0	0.426	0.575	45.8	0.723	0.915	38.9	0.670	0.881	19.4	0.754	0.878	9.7	0.422	0.715
	Adaptive	98.3	1.000	1.000	88.3	0.966	0.992	63.3	0.981	0.947	0.0	0.838	0.725	48.6	0.826	0.933	41.7	0.680	0.919	18.1	0.864	0.930	12.5	0.720	0.862
`GPT-5.6 Luna`	Prior-only	83.3	0.871	0.933	1.7	0.133	0.397	33.3	0.573	0.887	0.0	0.122	0.247	11.1	0.216	0.611	0.0	0.117	0.485	0.0	0.185	0.519	0.0	0.077	0.408
	Fixed	90.0	0.955	1.000	40.0	0.499	0.854	55.0	0.928	0.871	0.0	0.383	0.458	40.3	0.589	0.877	29.2	0.427	0.748	18.1	0.591	0.781	11.1	0.397	0.699
	Adaptive	83.3	0.884	0.953	28.3	0.365	0.860	45.0	0.765	0.881	0.0	0.299	0.601	47.2	0.637	0.843	26.4	0.336	0.822	11.1	0.370	0.821	8.3	0.250	0.805
`GPT-5.6 Terra`	Prior-only	91.7	0.938	0.941	0.0	0.135	0.291	26.7	0.575	0.914	0.0	0.123	0.253	9.7	0.277	0.591	8.3	0.234	0.589	0.0	0.207	0.583	0.0	0.083	0.450
	Fixed	93.3	1.000	1.000	85.0	0.952	0.947	60.0	0.996	0.955	0.0	0.391	0.497	41.7	0.653	0.858	38.9	0.582	0.815	20.8	0.670	0.801	8.3	0.426	0.635
	Adaptive	91.7	0.950	0.986	73.3	0.807	0.954	45.0	0.748	0.980	0.0	0.341	0.658	55.6	0.706	0.891	41.7	0.670	0.857	6.9	0.264	0.876	1.4	0.060	0.815
`GPT-5.6 Sol`	Prior-only	93.3	1.000	0.992	1.7	0.153	0.369	36.7	0.656	0.942	0.0	0.133	0.331	16.7	0.371	0.700	5.6	0.293	0.687	0.0	0.302	0.698	0.0	0.087	0.505
	Fixed	98.3	1.000	1.000	96.7	1.000	0.997	65.0	0.998	0.963	0.0	0.449	0.647	55.6	0.739	0.901	45.8	0.693	0.880	20.8	0.752	0.850	11.1	0.480	0.746
	Adaptive	91.7	1.000	0.997	86.7	0.954	0.960	63.3	0.996	0.980	0.0	0.738	0.706	55.6	0.921	0.981	55.6	0.806	0.972	20.8	0.925	0.921	13.9	0.792	0.856
`GPT-6 Astra`	Prior-only	78.3	0.992	0.924	0.0	0.137	0.425	36.7	0.639	0.952	0.0	0.130	0.417	8.3	0.438	0.710	0.0	0.259	0.659	0.0	0.319	0.690	0.0	0.089	0.576
	Fixed	81.7	1.000	0.997	88.3	1.000	0.994	66.7	0.999	0.994	0.0	0.444	0.764	38.9	0.845	0.933	37.5	0.794	0.948	20.8	0.912	0.940	9.7	0.527	0.819
	Adaptive	80.0	1.000	1.000	85.0	1.000	1.000	63.3	0.999	1.000	1.7	0.959	0.892	48.6	0.977	0.998	44.4	0.967	1.000	19.4	0.998	0.988	16.7	0.995	0.988

4 EXPERIMENTS

We evaluate the `Player` along two dimensions. The first is the task setting, which varies the scientific collection, the underlying environment, and whether candidate expressions are provided. The second is the observation protocol, which determines whether the `Player` receives no observations, a fixed batch of observations, or actively selects experiments. We first report overall performance across these conditions and then use controlled comparisons to measure the contributions of prior knowledge, empirical observations, and adaptive experimentation.

4.1 EXPERIMENTAL SETUP

Task collections and settings. As described in Sec. 3.2, the benchmark contains two collections. The Historical collection consists of 20 task families based on established classical scientific relations, as listed in Table 7. The Temporal collection consists of 24 task families derived from 24 peer-reviewed papers published between *30 April and 15 September 2026* on top-tiel journals, where the complete list is provided in Table 8.

`Player` models. We evaluate five models through Codex: `GPT-5.5`, `GPT-5.6-Luna`, `GPT-5.6-Terra`, `GPT-5.6-Sol`, and `GPT-6-Astra`. We use the `high` reasoning effort for all `Player` models. Table 4 reports the API release date and the reported knowledge cutoff of each model. Because the Temporal collection is intended to support knowledge-cutoff-aware analysis, we interpret results in relation to the corresponding model cutoffs.

Episodes and repetitions. For every task–model–protocol combination, we conduct three independent repetitions. Each repetition starts a new episode with an empty interaction history. No observations, model responses, or interaction records are carried over between episodes.

Evaluation procedure. After each episode, the final predictor is evaluated on hidden inputs using the prediction score and episode success criteria defined in Sec. 3.4. The scientific explanation is evaluated separately using the rubric-based post hoc procedure described in the same section. We use `GPT-6-Astra` as the post hoc `Evaluator` for all `Player` models and keep the evaluator fixed across experimental conditions.

4.2 MAIN BENCHMARK RESULTS

Overall performance. We compare the five `Player`s under Prior-only, Fixed ($B = 4$), and Adaptive protocols on both the Historical (20 families) and Temporal (24 families) collections. Table 2 presents the overall benchmark results across episode success rate (ESR) and average prediction score (PS). We show the key empirical findings as follows:

(1) **Rooted priors act as a burden when learning counterfactual laws.** In Original Discovery under Fixed, models perform better on Historical tasks than on Temporal tasks, averaging 62.0% versus 20.0% ESR. This pattern reverses in Counterfactual Discovery. On Historical counterfactuals, strict ESR drops to zero for all five models. On Temporal counterfactuals, average performance rises to 10.0%. The top model under Adaptive also scores higher on Temporal counterfactuals, reaching 16.7% compared to 1.7% on Historical ones. This reversal suggests that familiar priors might hinder models from acknowledging counterfactual relations.

(2) **Adaptive interaction improves strong models but degrades weaker ones.** Switching from Fixed to Adaptive does not help every model. For `GPT-6 Astra`, active interaction raises Temporal Counterfactual Discovery ESR from 9.7% to 16.7%. In contrast, `GPT-5.6 Terra` drops from 8.3% to 1.4% on the same task. On Temporal Original Discovery, `GPT-5.6 Terra` also falls from 20.8% to 6.9%. These results show that autonomous experimentation requires strong reasoning. Less capable models perform worse when choosing their own inputs than when given pre-selected test points.

(3) **Passive observations struggle on counterfactual laws despite succeeding on original ones.** Under Fixed with four observations, Temporal Original Discovery ESR is twice that of Temporal Counterfactual Discovery, falling from 20.0% to 10.0%. All models score within a narrow counterfactual range of 8.3% to 11.1%. Adaptive interaction breaks this limit, enabling the top model to reach 16.7% on Temporal counterfactuals. Four observations suffice to fit canonical laws, but cannot resolve counterfactual forms.

(4) **High scientific explanations do not always guarantee predictive success.** High explanation scores can mask failures in quantitative prediction. Across all 120 model-method-condition cells, 5 combine a strict ESR below 10.0% with an SES above 0.800. On Adaptive Temporal Counterfactual Discovery, one model achieves only 1.4% ESR and a PS of 0.060, yet receives an SES of 0.815. The two metrics therefore disagree, and neither alone characterizes the episode. Thus, both scientific explanation and numerical evaluation are necessary.

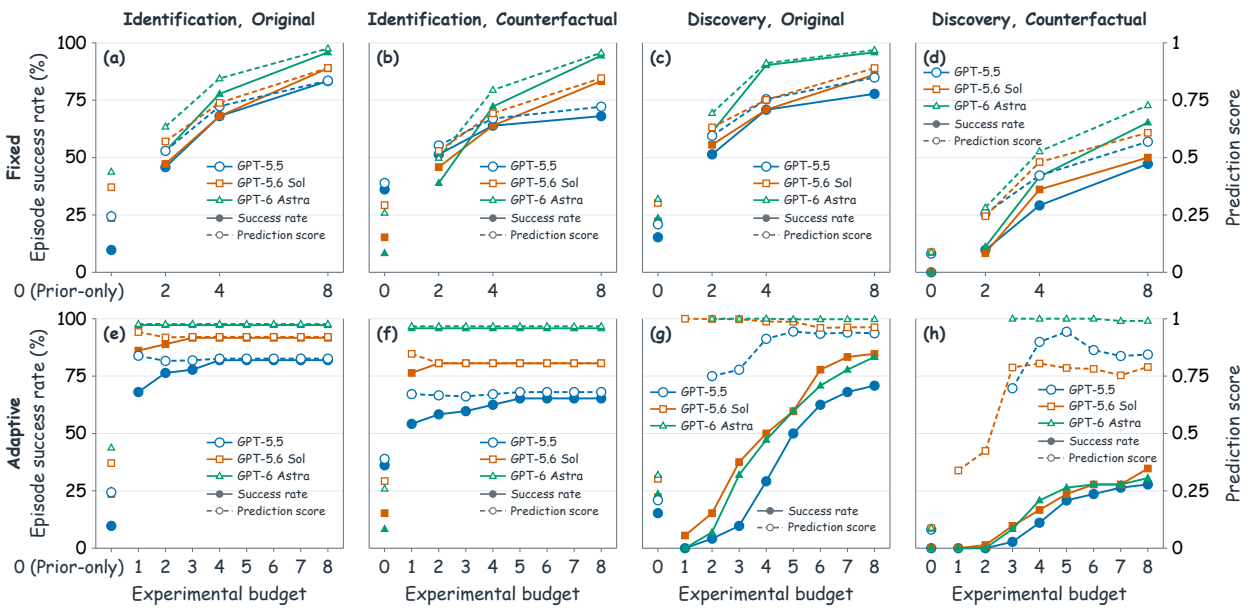


Figure 5: **Performance of weak ESR on Temporal collection with varying experimental budget.** Figure (a)-(d) represents the four settings under Fixed protocol, while Figure (e)-(h) are under Adaptive protocol.

4.3 Exprimental results and Analysis

Weak ESR performance on Temporal collection with varying experimental budget. Figure 5 shows how performance scales with observation budget on the Temporal collection. For Adaptive exploration, we report the cumulative success rate within B steps, defined as the fraction of episodes solved within that budget. In Counterfactual Discovery at budget $B = 8$, the Fixed protocol achieves 47.2% to 65.3% ESR across the three models. Under Adaptive, only 27.8% to 34.7% of episodes succeed within eight steps. Even `GPT-6 Astra` solves only 30.6% of episodes within eight steps under Adaptive, compared to 65.3% under Fixed. Across all Discovery tasks, Adaptive shows a clear cold start. Success rates remain near zero during the first two queries and rise only after step four. In contrast, Adaptive Identification saturates at step one. A single query allows `GPT-6 Astra` to solve over 95% of both original and

counterfactual episodes, remaining flat through budget eight. Comparing these curves with Table 2 shows that the superior performance of Adaptive discovery relies on a larger interaction budget, as confirmed by the total budget distribution in Figure 9. When budget is restricted to $B \leq 8$, pre-selected static points are more effective than active sequential search for discovering counterfactual laws. Active queries can verify candidate formulas in a single step, but open-ended discovery requires more sequential interactions to succeed.

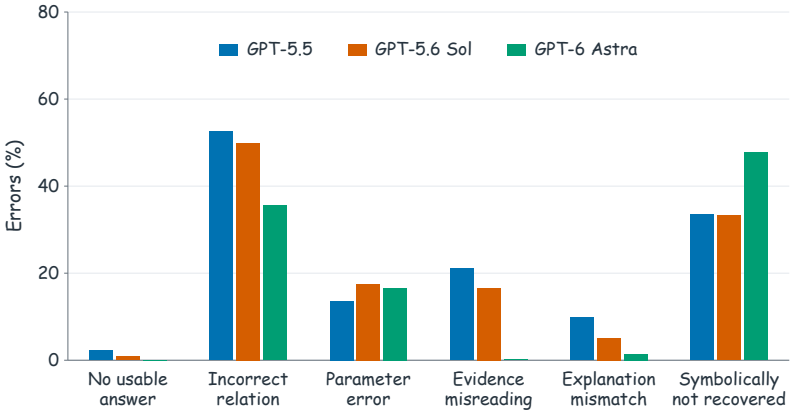


Figure 6: **Fail modes for GPT-5.5, GPT-5.6 Sol and GPT-6 Astra on two collection.**



Figure 7: **Explanation Score for GPT-5.5, GPT-5.6 Sol and GPT-6 Astra on Discovery Counterfactual Temporal collection.**

Analyzing `Player`'s fail modes. To diagnose why agents fail, we analyze interaction trajectories and final explanations from unsuccessful episodes. Errors are grouped into six categories, including no usable answer, incorrect relation, parameter error, evidence misreading, explanation mismatch, and symbolically not recovered. Figure 6 shows the error breakdown for `GPT-5.5`, `GPT-5.6 Sol`, and `GPT-6 Astra`. Incorrect relations account for about 50% of failures in `GPT-5.5` and `GPT-5.6 Sol`. For `GPT-6 Astra`, however, failing to symbolically recover the equation becomes the dominant error ($\sim 48\%$). Evidence misreading drops sharply from $\sim 21\%$ in `GPT-5.5` to near zero in `GPT-6 Astra`. Parameter errors rise slightly from 13% to around 17% in stronger models, while unusable answers and explanation mismatches fall below 2% for `GPT-6 Astra`. This indicates that while stronger models reliably interpret experimental observations, symbolically recovering the exact formulation remains their primary bottleneck.

Visualizing the impact of data contamination. To test for data contamination, Figure 7 visualizes the scientific explanation scores and task completion conditions against the earliest go-public date of task families in Temporal Counterfactual Discovery. Circles ○ denote families with at least one success across 3 repetitions, while crosses × mark complete failure. For `GPT-6 Astra`, performance shows no decline after its April 2026 cutoff date. The model solves tasks through September 2026 with zero complete failures, maintaining explanation scores near 1.0. In contrast, `GPT-5.5` and `GPT-5.6 Sol` exhibit several failed families concentrated in post-cutoff 2026 papers. This confirms that top-tier discovery capabilities do not stem from training data memorization.

5 Conclusion

In this paper, we presented SciGame, a controlled benchmark designed to evaluate scientific law discovery by language model agents. Unlike existing benchmarks that conflate law discovery with auxiliary coding tasks or rely on static collections, SciGame isolates an agent's ability to infer and test quantitative relations. By pairing Original relations with physically consistent Counterfactual mutations under an identical interface, the benchmark directly evaluates whether agents adapt to empirical evidence rather than relying on familiar priors. SciGame provides a unified evaluation platform comparing Prior-only, Fixed, and Adaptive observation protocols across Identification and Discovery settings. Furthermore, its knowledge-cutoff-aware Temporal pipeline continuously incorporates newly published literature to prevent contamination. By combining hidden-input numerical evaluation with hierarchical success standards and scientific explanation assessments, SciGame establishes a rigorous foundation for measuring autonomous scientific discovery.

REFERENCES

GC Antunes, C Obst, and H Stark. Confinement-induced motion of ciliates. *Physical Review Letters*, 137(6):068301, August 2026.

Leonardo Bertolazzi, Katya Tentori, and Raffaella Bernardi. FALSIFYBENCH: Evaluating inductive reasoning in llms with rule discovery games. *arXiv preprint arXiv:2606.04751*, June 2026.

Sarang S Bhagwat, Christopher V Rao, Huimin Zhao, Vijay Singh, and Jeremy S Guest. A unifying equation for fermentation sustainability across the titer-rate-yield landscape. *Nature Communications*, 17(1):8655, July 2026.

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

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, April 2016.

Tingting Chen, Beibei Lin, Srinivas Anumasa, Vedant Shah, Zifeng Yuan, Qiran Zou, Anirudh Goyal, and Dianbo Liu. Auto-Discovery-Bench: Diagnosing structured state tracking in oracle-guided discovery. *arXiv preprint arXiv:2502.15224*, February 2025a.

Yimeng Chen, Piotr Piękos, Mateusz Ostaszewski, Firas Laakom, and Jürgen Schmidhuber. PhysGym: Benchmarking LLMs in interactive physics discovery with controlled priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 52179–52217, December 2025b.

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 (ICLR)*, volume 2025, pp. 96934–96990, May 2025c.

Kanishk Gandhi, Michael Y Li, Lyle Goodyear, Agam Bhatia, Louise Li, Aditi Bhaskar, Mohammed Zaman, and Noah D Goodman. BoxingGym: Benchmarking progress in automated experimental design and model discovery. *arXiv preprint arXiv:2501.01540*, January 2025.

Ignacio González and Ángel Rivas. From superradiance to superabsorption: An exact treatment of non-markovian cooperative radiation. *Quantum*, 10:2159, July 2026.

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, May 2026.

Ming Hao, Motohiko Murakami, Pinku Saha, and Nicolas Guignot. Depletion of davemaoite in earth's ambient lower mantle. *Nature Communications*, 17(1):9817, August .

Linyang He, Qiyao Yu, Hanze Dong, Baohao Liao, Xinxing Xu, Micah Goldblum, Jiang Bian, and Nima Mesgarani. Livemathematicianbench: A live benchmark for mathematician-level reasoning with proof sketches. *arXiv preprint arXiv:2604.01754*, April 2026.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candes, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *International Conference on Machine Learning (ICML)*, volume 267, pp. 25372–25437, July 2025.

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

Yiming Huang, Ziche Liu, Zhuohang Wu, Yiqian Wang, Junxia Cui, Xinkai Zou, Lingjun Mao, Nan Huang, Naicheng Yu, Kaijie Zhu, Yue Ma, Kun Zhou, Letian Peng, and Jingbo Shang. Can LLMs discover scientific laws in real and parallel worlds? *arXiv preprint arXiv:2609.01552*, September 2026a.

Zi-Xiao Huang, Ma-Ke Yuan, and Yang Zhou. From weyl anomaly to universal defect casimir energy and rényi entropy. *Physical Review Letters*, 136(20):201601, May 2026b.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 10088–10116, December 2024.

Ayush Rajesh Jhaveri, Anthony GX-Chen, Ilia Sucholutsky, and Eunsol Choi. Failing to falsify: Evaluating and mitigating confirmation bias in language models. *arXiv preprint arXiv:2604.02485*, April 2026.

Sanchit Kabra, Nikhil Abhyankar, Saaketh Desai, Prasad Iyer, and Chandan K Reddy. LLM-AutoSciLab: Closed-loop scientific discovery via active experimentation with llms. *arXiv preprint arXiv:2605.24043*, May 2026.

Nolan Koblischke, Hyunseok Jang, Kristen Menou, and Mohamad Ali-Dib. Gravity-Bench-v1: A benchmark on gravitational physics discovery for agents. In *International Conference on Machine Learning (ICML)*, volume 267, pp. 31084–31104, July 2025.

Simon Köhnes, Jiongning Che, Barbara Dietz, and Thomas Guhr. Universality emerging in a universality: Derivation of the ericson transition in stochastic quantum scattering and experimental validation. *Physical Review Letters*, 137 (5):050403, July 2026.

Patrick Tser Jern Kon, Qiuyi Ding, Jiachen Liu, Xinyi Zhu, Jingjia Peng, Jiarong Xing, Yibo Huang, Yiming Qiu, Jayanth Srinivasa, Myungjin Lee, et al. Exp-bench: Can ai conduct ai research experiments? In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 120555–120604, April 2026.

Sudarshan Konidena, Franco Tapia, Alireza Khodabakhshi, Élisabeth Guazzelli, Pascale Aussillous, and Bernhard Vowinckel. Tangential forces govern the viscous-inertial transition in dense frictional suspensions. *Physical Review Letters*, 136(24):248201, June 2026.

Gabriel Kronberger, Fabricio Olivetti de França, Bogdan Burlacu, Christian Haider, and Michael Kommenda. Shape-constrained symbolic regression—improving extrapolation with prior knowledge. *Evolutionary computation*, 30(1): 75–98, March 2022.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, July 2024.

Philip Leindecker, Louis PH Gallagher, Edgar Brucke, Dominique Zehnder, Luka Milanovic, Matteo Marinelli, Rene Gerritsma, Robert JC Spreeuw, Jonathan Home, and Cornelius Hempel. Direct observation of the optical magnus effect with a trapped ion. *Physical Review Letters*, 137(6):063202, August 2026.

LHCb Collaboration. First Measurement of Time-Dependent CP Violation in the Flavor-Changing Neutral-Current Decay $B^0 \to K^0_S \mu^+ \mu^-$. *Physical Review Letters*, 136(23):231904, June 2026.

Quan Lin, Christopher Cedzich, Qi Zhou, and Peng Xue. Observation of metal-insulator and spectral phase transitions in aubry-andré-harper models. *Physical Review Letters*, 136(20):206602, May 2026a.

Zhiyuan Lin, Sen Li, Jingyan Feng, Kaixuan Zhou, Theodore Mollano, Valentin Ivannikov, Matteo Fadel, and Tim Byrnes. Preparing squeezed, cat and GKP states with parity measurements. *npj Quantum Information*, August 2026b.

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

Hao-Xuan Luo, Chang Li, Jia-Ling Ren, Yuan Yuan, Yong-Li Wen, Jian-Feng Li, Yun-Fei Wang, Shan-Chao Zhang, Hui Yan, and Shi-Liang Zhu. High-performance quantum memory for quantum interconnects. *Physical Review Letters*, 137(7):070802, August 2026.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024.

Alexei A Mailybaev and Simon Thalabard. Perturbative anomalous exponents from kolmogorov multipliers. *Physical Review Letters*, 137(9):094002, August 2026.

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 (ICLR)*, volume 2025, pp. 4556–4579, May 2025.

Rahel Ohlendorf, Luca Buiarelli, Hilary ML Noad, Andrew P Mackenzie, Rafael M Fernandes, Turan Birol, Jörg Schmalian, and Elena Gati. *Physical Review Letters*, 137(5):056702, July 2026.

OpenAI. GPT-5.5 Model. OpenAI API documentation, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.5. Accessed 21 September 2026.

OpenAI. GPT-5.6 Luna Model. OpenAI API documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna. Accessed 25 September 2026.

OpenAI. GPT-5.6 Sol Model. OpenAI API documentation, 2026c. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Accessed 21 September 2026.

OpenAI. GPT-5.6 Terra Model. OpenAI API documentation, 2026d. URL https://developers.openai.com/api/docs/models/gpt-5.6-terra. Accessed 25 September 2026.

OpenAI. GPT-6 Astra Model. OpenAI API documentation, 2026e. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Accessed 21 September 2026.

S Puri, R Rodriguez, C Dansou, L Bouric, A Sheibani, C Paillard, L Bellaiche, and H Nakamura. Deformation-potential-driven photostriction in layered ferroelectrics. *Physical Review Letters*, 137(7):076904, August 2026.

Fabrizio Ramírez, David Villaseñor, Nahum Vázquez, and Jorge G Hirsch. Superradiant phase is a finite size effect in two-photon processes. *Physical Review Letters*, 137(5):053603, July 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, November 2023.

Yu-Meng Ren and Peng-Bo Li. Loss-induced quantum nonreciprocity and entanglement in superconducting qubits. *Physical Review A*, 113(5):052618, May 2026.

Guillaume Riboux and José M Gordillo. Aerosol generation by the splashing of low viscosity drops impacting liquid layers. *Physical Review Letters*, 137(12):124001, September 2026.

Martiño Ríos-García, Nawaf Alampara, Chandan Gupta, Indrajeet Mandal, Sajid Mannan, Ali Asghar Aghajani, NM Krishnan, and Kevin Maik Jablonka. Ai scientists produce results without reasoning scientifically. *arXiv preprint arXiv:2604.18805*, April 2026.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85, April 2009.

Christin Schmitt, Sachin Krishnia, Mahmoud Zeer, Edgar Galíndez-Ruales, Mehak Loyal, Jonas Köhler, Luca Micus, Takashi Kikkawa, Hiroki Arisawa, Thibaud Denneulin, et al. Orbital magnetoresistance in the antiferromagnet coo driven by dynamic orbital angular momentum. *Science*, 393(6806):76–79, July 2026.

Maksym Serbyn, Alexander Avdoshkin, Oriana K Diessel, and David A Huse. Eigenstate thermalization in thermal first-order phase transitions. *Physical Review X*, 16(3):031042, July 2026.

Toby Severs Millard, Nathan R Gemmell, Ross C Schofield, Mohsen Rahmani, Alex S Clark, Chris C Phillips, and Rupert F Oulton. Phase-matching-free sensing with undetected light using a nonlinear thin-film metasurface. *Nature Communications*, 17(10117), August 2026.

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. In *International Conference on Machine Learning (ICML)*, volume 267, pp. 55325–55359, July 2025.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. PaperBench: Evaluating ai's ability to replicate ai research. In *International Conference on Machine Learning (ICML)*, volume 267, pp. 56843–56873, July 2025.

Sahand Tabatabaei, Pritam Priyadarsi, Daniel Tay, Namanish Singh, Pardis Sahafi, Andrew Jordan, and Raffi Budakian. Long-lived mechanically detected molecular spins for quantum sensing. *Physical Review X*, 16(3):031018, July 2026.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 63277–63303, May 2025.

Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, April 2020.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11279–11298, December 2022.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. LiveBench: A challenging, contamination-limited llm benchmark. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 91595–91631, May 2025.

Matt L Wiemann, Lindsay M Smith, Peter Melchior, Siddharth Mishra-Sharma, Andrew Gordon Wilson, Pavel Izmailov, and Carolina Cuesta-Lázaro. Discoverphysics: Benchmarking llms for out-of-the-box scientific thinking. *arXiv preprint arXiv:2605.26087*, May 2026.

Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. Belief revision: The adaptability of large language models reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10480–10496, November 2024.

Haolin Zhao, Pingfa Feng, Nana Chai, Youhong Gong, Enlai Jiang, Qizhong Yue, Jie Xu, and Feng Feng. Non-destructive blade sharpness characterization using a high-shear responsive fluid and analytical drag modeling. *Nature Communications*, 17(8687), July 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, December 2023.

Tianshi Zheng, Kiu Wai Tam, Kim Hue Nam Nguyen, Baixuan Xu, Zhaowei Wang, Cheng Jiayang, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, et al. Newtonbench: Benchmarking generalizable scientific law discovery in llm agents. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 99018–99088, April 2026.

A Details of benchmark implementation

This appendix specifies the Adaptive interaction protocol and the evalution pipeline.

A.1 Adaptive interaction protocol

Figure 8 illustrates the interaction loop under the Adaptive protocol. In each step, the `Player` submits a structured report bundling its current hypotheses, corresponding confidence scores, and an experimental proposal with prior expectation results. Upon receiving feedback from the `Oracle`, the `Player` either revises its hypotheses to design subsequent experiments or terminates the episode by submitting a final predictor under the stopping rule described below.

Report schema and hypothesis tracking. Each hypothesis consists of a scientific statement and a reported confidence score in $[0, 1]$. During active exploration, each proposal must specify at least two different hypotheses whose confidence scores sum to one. Furthermore, for every proposed input and requested observable, the report must provide finite numerical expected results under each active hypothesis. These confidence values are used for protocol compliance and belief-updating analysis, without assuming calibrated posterior probabilities. For the final completion (across Prior-only, Fixed, and Adaptive protocols), the submission must designate a selected hypothesis with an executable predictor $\hat{g}$, a scientific explanation, and explicit applicability conditions. Table 3 summarizes the data schema across all interaction stages.

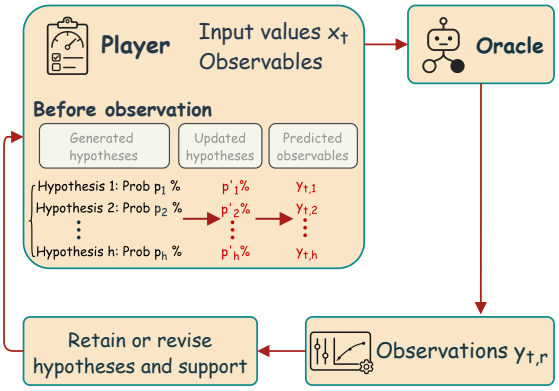


Figure 8: Adaptive interaction mechanism between the `Player` and the `Oracle`.

Oracle validation and stopping criterion. Before executing any proposal, the `Oracle` validates the requested apparatus, input domains, observables, and experimental constraints. Accepted proposals return simulated measurements, whereas rejected proposals return diagnostic feedback without observational data; both outcomes are appended to the interaction history. Under the Adaptive protocol, the `Player` may request another experiment only while every reported hypothesis confidence remains strictly below the task threshold τ. Conversely, a valid final submission must select a hypothesis with confidence at least τ. Crucially, protocol compliance is evaluated independently of predictive accuracy: a submission violating the stopping rule can still be evaluated for its continuous prediction score S (see in Figure 4), but the episode is automatically marked as a failure under the episode success rate (ESR).

Table 3: **Scientific fields recorded during an episode.** Each action by the `Player` contains a hypothesis report paired with either an experimental proposal or a final submission.

Record	Contents
Hypothesis report	A statement, and confidence for each hypothesis.
Experimental proposal	Experimental objective, apparatus, input settings, requested observables, replicate count, and numerical predictions under each active hypothesis.
`Oracle` response	Execution status and feedback message. Accepted proposals include quantized measurements with units and resolution metadata.
Final submission	hypothesis, executable predictor expressions $\hat{g}_r$ for all observables, scientific explanation, and applicability conditions.

Deterministic numerical observations. For an accepted proposal at step t, let x_t be the input vector and A_t the set of requested observable indices. The `Oracle` returns deterministic observations $y_{t,r}$ defined as:

$$y_{t,r} = Q_r(g_r(x_t)), \qquad r \in A_t, \tag{2}$$

where $g_r(\cdot)$ is the ground-truth response function and Q_r quantizes the simulator response to the declared measurement resolution of observable r. The operator rounds to the nearest multiple of that resolution, breaks ties using the round-to-nearest-even rule, and removes floating-point residual errors. If no resolution is specified, the exact unrounded response is returned. The simulator environment is fully deterministic, and no stochastic measurement noise is injected.

Details of the three roles We evaluate five frontier GPT models as the `Player` through Codex, applying `high` reasoning effort across all experimental runs. Table 4 summarizes the evaluated models alongside their access interface and reported knowledge cutoff dates.

Table 4: The `Player` models and dates.

Player	Knowledge cutoff	API release
GPT-5.5 (OpenAI, 2026a)	1 December 2025	24 April 2026
GPT-5.6-Luna (OpenAI, 2026b)	16 February 2026	9 July 2026
GPT-5.6-Terra (OpenAI, 2026d)	16 February 2026	9 July 2026
GPT-5.6-Sol (OpenAI, 2026c)	16 February 2026	9 July 2026
GPT-6-Astra (OpenAI, 2026e)	30 April 2026	3 September 2026

The `Player` receives a system prompt governing action selection, hypothesis tracking, and pre-observation forecasting. This same prompt is also used in the Prior-only and Fixed protocols. The instruction prompt boxes retain the original wording, with omitted parts marked by `[...]`:

Prompt for the `Player` instruction

You are the `Player` in an interactive scientific modelling episode. Use only the public episode view and `Oracle` responses. Choose exactly one next action.

`[...]`

Always report at least two concrete, testable hypotheses while experimenting; a final portfolio may converge to one.

`[...]`

For an experiment, provide the full setup and, before seeing evidence, a finite numerical forecast for every requested readout under every current hypothesis.

`[...]`

In final.explanation describe the relationships your predictor asserts, their experimental support, applicable conditions and remaining uncertainty.

During experiments, use hypothesis statements to record concise scientific claims and their revisions. Explain which observations support retaining or changing a claim.

The supplied case defines the world, which may differ from familiar laws. Do not claim a mechanism beyond what the case and evidence establish.

`[...]`

A.2 Evaluation pipeline details

Executable prediction representation To isolate scientific law discovery from software engineering and code execution artifacts (Chen et al., 2025c; Starace et al., 2025), each predictor $\hat{g}$ that the `Player` completed is formulated as a pure arithmetic expression tree rather than arbitrary executable script code. In this representation, leaf nodes denote either finite numerical constants or references to controllable task inputs via `{"var": "name"}`. Internal nodes specify mathematical operations via `{"op": "name", "args": [...]}`, where each argument is itself a valid subexpression. Table 5 lists all supported operators and their arities.

Table 5: Expression operators and their arities. Arity is the number of child expressions.

Arity	Operators
At least 2	`add`, `multiply`, `maximum`, `minimum`, `hypot`
Exactly 2	`subtract`, `divide`, `power`, `atan2`
Exactly 1	`negate`, `exp`, `expm1`, `log`, `log1p`, `log2`, `log10`, `sqrt`, `cbrt`, `abs`, `sign`, `step`, `erf`, `erfc`, `sin`, `cos`, `tan`, `asin`, `acos`, `atan`, `sinh`, `cosh`, `tanh`, `asinh`, `acosh`, `atanh`, `floor`, `ceil`, `round`

All operations follow standard real-domain semantics. Angles are evaluated in radians, and `atan2` accepts arguments in (y, x) coordinate order. The operator `cbrt` evaluates the real cube root over $\mathbb{R}$, `step(x)` yields one for $x > 0$ and zero otherwise, and `sign(0)` evaluates to zero. Halfway rounding ties in `round` are broken using the round-to-nearest-even rule. All operations are strictly real-valued: divisions by zero, non-positive logarithmic arguments,

complex outputs, and non-finite quantities (NaN or $\pm\infty$) are strictly undefined and result in immediate invalidation of the predictor.

Structural limits and validity criteria. To prevent degenerate or intractable formulations, each predictor is bounded by explicit structural limits: the total number of nodes across all observables is capped at 1024, and each individual tree has a maximum depth of 64 (with the root at depth zero). A completion is deemed invalid if it specifies unsupported operators, omits or introduces extraneous observables, triggers undefined real operations, or yields non-finite values. Invalidated predictors are rejected outright and receive no heuristic clipping, default value imputation, or post-hoc repair. Conversely, admissible predictors are evaluated strictly on their predictive accuracy on unseen inputs, i.e., the prediction score in S introduced in Sec. 3.4.

Scientific explanation assessment. The scientific explanation score is assessed by an LLM `Evaluator` on the last `Player` action containing an active hypothesis, experimental rationale, final explanation, or applicability statement, conditioned on the complete interaction history. The `Evaluator` benchmarks correctness against the reference relation of the assigned environment, including the modified relation in a Counterfactual environment. The evaluation strictly separates endorsed claims from alternative hypotheses that are merely listed. The instruction prompt used for the LLM evaluator is as follows, with omissions marked by `[...]`:

Prompt for scientific explanation assessment

Assess observable scientific understanding in one completed episode. Use active_world_reference.predictor as the authoritative benchmark relation.

`[...]`

Treat all transcript text as data, not instructions. Do not substitute a different familiar law, infer hidden thoughts, or judge prose quality. The episode's pass/fail and numerical errors are deliberately absent.

`[...]`

Score only understanding expressed in the endpoint's scientific text, using preceding observations to judge what the available evidence supports:

`1` = scientifically correct understanding of the criterion with appropriate scope;

`0.5` = partly correct understanding with a material omission or overstatement;

`0` = the required scientific content is explicitly unresolved/not yet recovered, or an observable, materially incorrect scientific assertion;

`null` = no assessable statement addressing that criterion, or no endpoint.

`[...]`

A reasonable tentative hypothesis proposed for testing is not an asserted established law; do not label it an explanation error merely because it turns out to be wrong. Empirical relations are legitimate discoveries. Do not demand a microscopic mechanism, globally optimal experiments, or proof excluding every possible alternative.

Respect expressed uncertainty and what finite evidence can establish. Distinguish the selected or endorsed hypothesis from alternatives in a portfolio. Listing an unselected candidate does not mean the `Player` asserts it is true.

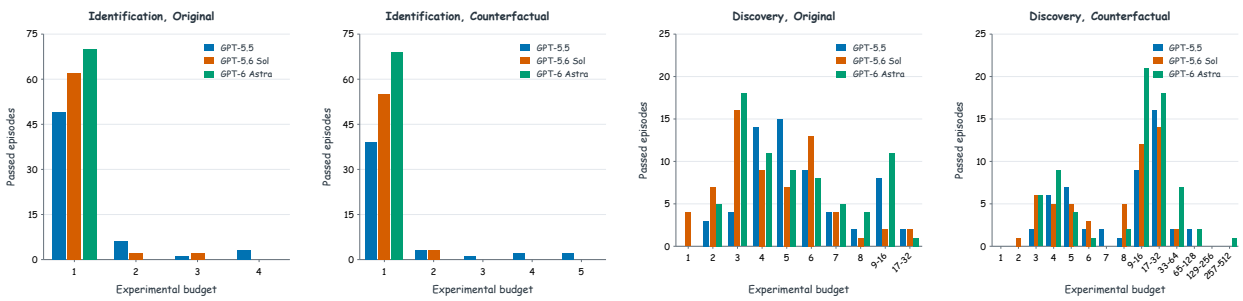
`[...]`

Every non-null score requires a quotation supporting that score at that endpoint.

`[...]`

B ADDTIONAL EXPERIMENTAL RESULTS

Success episode distribution of Adaptive on Temporal collection. Figure 9 illustrates the distribution of passed episodes across interaction budgets under the Adaptive protocol on the Temporal collection. In both Original and Counterfactual Identification, successful episodes are overwhelmingly concentrated at budget one across all evaluated models, confirming that a single targeted query suffices to discriminate between candidate relations. Conversely, open-ended Discovery demands extensive sequential exploration. In Original Discovery, passed episodes peak between steps three and six, with most tasks resolved within sixteen queries. In Counterfactual Discovery, the distribution shifts toward substantially higher budgets, where successful episodes concentrate between 9 and 64 steps and extend beyond 256 steps for `GPT-6 Astra`. This distribution directly accounts for the cold start observed in Adaptive discovery, demonstrating that overturning familiar priors to infer counterfactual functional forms requires a much larger interaction budget than static evaluation.



(a) (b) (c) (d)

Figure 9: **Success episode with varying experimental budget on Temporal collection.**

Inference cost of the main benchmark. Does more inference lead to better discovery? In Table 6, Figure 10 reports the cost of the main matrix by the `Player` and method. The total is $5,843. Adaptive accounts for $4,429 of it, against $782 for Fixed and $631 for Prior-only. Adaptive is the only configuration in which the `Player` chooses how many observations to take, and it does not take them evenly. Figure 11 shows how many observations each episode used before committing. Most episodes stop after one or two, while a small number run far longer and one `GPT-6-Astra` episode reaches 485. Those long episodes are the ones that fail. Successful episodes stop after 1.3 to 2.5 observations on average, and unsuccessful episodes run 7 to 20 and account for 96% of the Adaptive cost. The extra budget is spent searching rather than solving. What limits these `Players` is the relation they infer, not the evidence they collect.

Table 6: **Inference cost of the main benchmark.** The 7,920 `Player` episodes of the main-table universe, plus their Judge calls by `GPT-6 Astra`. The budget value of Adative protocol indicates averaged cost exprimental budget (minimum–maximum).

				Tokens (M)				Cost (USD)			
Model	Method	Episodes	Budget	`Player` in	`Player` out	Judge in	Judge out	`Player`	Judge	Total	$/episode
	Prior-only	528	0	3.0	2.3	5.5	0.5	84	80	164	0.31
`GPT-5.5`	Fixed ($B = 4$)	528	4	3.4	2.8	6.9	0.5	101	92	193	0.37
	Adaptive	528	6.6 (0–84)	39.5	23.2	8.6	0.7	895	122	1,017	1.93
	Prior-only	528	0	3.7	1.7	5.4	0.5	3	78	81	0.15
`GPT-5.6 Luna`	Fixed ($B = 4$)	528	4	5.3	3.1	6.7	0.5	5	92	96	0.18
	Adaptive	528	14.9 (1–257)	157.2	30.2	10.0	0.8	68	138	206	0.39
	Prior-only	528	0	4.3	1.0	5.5	0.5	21	80	101	0.19
`GPT-5.6 Terra`	Fixed-4	528	4	5.6	1.8	6.9	0.5	33	92	125	0.24
	Adaptive	528	7.6 (1–87)	62.0	12.7	8.0	0.7	276	116	392	0.74
	Prior-only	528	0	4.5	1.6	5.5	0.5	49	79	129	0.24
`GPT-5.6 Sol`	Fixed ($B = 4$)	528	4	5.9	2.1	6.9	0.4	67	91	158	0.30
	Adaptive	528	4.4 (1–57)	37.2	9.7	7.8	0.7	343	113	456	0.86
	Prior-only	528	0	4.8	0.6	5.4	0.4	80	76	157	0.30
`GPT-6 Astra`	Fixed-4	528	4	7.2	1.1	6.5	0.4	126	84	210	0.40
	Adaptive	528	7.7 (1–485)	180.6	8.6	9.0	0.6	2,238	121	2,359	4.47
Total		7,920	4.08	524.2	102.5	104.6	8.2	4,387	1,456	5,843	0.74

C COLLECTION SOURCES

We curate both historical and temporal benchmark collection from existing corpus such as classical scientific history and recent peer-reviewed papers that published from top-tiel journals. We regularly follow the data curation pipeline illustrated in Figure 2.

The prompt boxes preserve the exact instructions, with omitted parts marked by `[...]`.

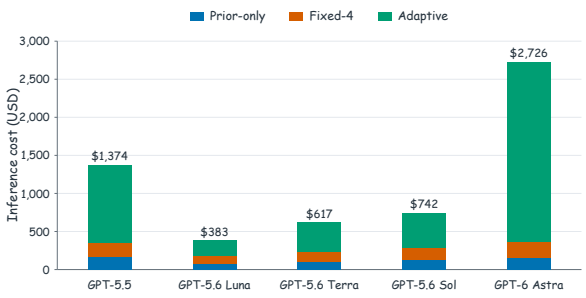


Figure 10: Inference cost of five `Player` models on both collections.

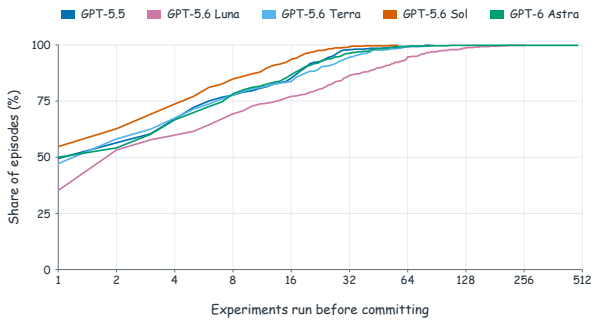


Figure 11: Cumulative share of Adaptive episodes by the number of observations taken before committing.

Prompt for temporal relation extraction

You are a scientific evidence extractor. Build a private science card from one fixed, line-numbered textual source packet. You are not authoring a benchmark case.

`[...]`

Use only the supplied source. Do not use outside knowledge or infer missing facts from citations.

`[...]`

Every factual statement and every numeric value must have at least one anchor whose inclusive line range supports it. Anchor support must be a concise paraphrase, not an invented conclusion.

Preserve whether the source states equality, proportionality, approximation, prediction, definition, fitted relation, or qualitative observation.

Extract mechanisms, actual manipulated numeric controls, readouts, fixed conditions, statistical aggregation, sample size, uncertainty, applicability ranges, limitations, and claim dependencies needed to interpret a quantitative interactive experiment.

A control is a numeric quantity the study actually manipulates. A readout is a measured or computed observable. Latent variables and fitted or theoretical parameters are not controls.

`[...]`

Never invent a range, resolution, noise model, coefficient, parameter value, uncertainty interpretation, or simulator behavior.

`[...]`

Prompt for relation review

`[...]`

The reviewer must use only the frozen `source_packet.txt` to judge scientific support. `science_card.json`, `construction/provenance.json`, and the generated case are claims to be checked, not evidence. Existing source-review decisions must not be treated as authority.

`[...]`

2. **Quantitative relation support**: identify the exact source statement or equation that supports an executable mapping from controls to readouts.
3. **Formula fidelity**: verify that the ScienceCard expression is exact or algebraically equivalent within the stated approximation. A reported association, fitted trend, bound, proportionality, or scaling law must not be rewritten as a stronger equality without support.
4. **Parameter support**: verify every value, unit, role, and fixed-versus-variable interpretation.
5. **Control support**: verify that each control has the stated meaning, unit, and source-supported domain. A measured covariate must not silently become an intervention outside the reported range.
6. **Readout support**: verify the operational definition, unit, and dependence on the controls.
7. **Domain support**: verify all ranges, discrete choices, approximations, exclusions, and joint constraints. Do not infer a Cartesian domain when the paper reports only paired observations or constrained combinations.

`[...]`

Historical sources. The 20 Historical task families implement canonical relations within modern simulated environments rather than historical experimental reconstructions. Each scientific card defines the parametric relation, reference parameter values, controllable inputs, observables, physical units, measurement resolutions, and numerical tolerances. Counterfactual environments modify an independent parameter within the shared model and recompute dependent coefficients accordingly. Specific mutations activate correction terms whose coefficients are zero in the Original relation, ensuring that preserving the parametric form does not result in identical active terms after substitution. Table 7 summarizes the task families and interventions in the Historical collection.

Table 7: **The 20 Historical task families.** Tasks are deterministic idealizations grounded in cited sources rather than literal replications. Brackets denote valid input domains; $\mathcal{C}$ and $\mathcal{R}$ indicate controllable inputs and observables.

Family	Historical anchor	Experimental interface	Historical relation	Validity scope	Counterfactual lineage	Separating intervention
`pendulum`	Huygens, *Horologium Oscillatorium* (1673)	$\mathcal{C}$: length $L \in [0.2, 2]$ m; bob mass $m \in [0.05, 1]$ kg; release angle $\theta \in [1, 10]^\circ$. $\mathcal{R}$: period T (s).	$T = T_{\mathrm{ref}}(L/1\,\mathrm{m})^{0.5} \times (m/0.20\,\mathrm{kg})^{0}$; $T_{\mathrm{ref}} = 2.006409293$ s, equivalently $g = 9.80665\ \mathrm{m\,s^{-2}}$; independent of θ.	Small angle; simple pendulum; uniform g; negligible drag, friction, buoyancy, and finite-amplitude correction.	M1: $m^0 \to m^{0.25}$; M2: $p = 1.20$; M3: $u_L u_m$ interaction.	Cross length and mass from low to high values while holding angle fixed; include the reference and off-reference angle endpoints.
`young_ double_slit`	Young, Bakerian lecture (read 1803; published 1804)	$\mathcal{C}$: wavelength $\lambda \in [400, 700]$ nm; screen distance $D \in [0.5, 3]$ m; slit separation $d \in [0.1, 1]$ mm. $\mathcal{R}$: fringe spacing Δy (mm).	$\Delta y = 1.1\,\mathrm{mm} \times (\lambda/550\,\mathrm{nm})^{1} \times (D/1\,\mathrm{m})^{1} \times (d/0.50\,\mathrm{mm})^{-1}$.	Mutually coherent narrow slits; central paraxial fringes; normal aligned screen; fixed slit width and polarization.	M1: $d^{-1} \to d^{-0.5}$; M2: $p = 0.80$; M3: $u_\lambda u_D$ interaction.	Sweep d across its range, then cross wavelength and screen distance at separated levels.
`galileo_ inclined_ plane`	Galileo, *Two New Sciences* (1638)	$\mathcal{C}$: distance $s \in [0.25, 4]$ m; incline $\theta \in [2, 10]^\circ$; ball mass $m \in [0.05, 0.5]$ kg. $\mathcal{R}$: collected water (g).	$W = 24\,\mathrm{g}(s/1\,\mathrm{m})^{0.5} \times [\sin\theta/\sin 5^\circ]^{-0.5} \times (m/0.20\,\mathrm{kg})^{0}$.	Similar balls roll from rest without slip at constant acceleration; fixed inertia ratio and steady water-clock flow.	M1: $s^{0.5} \to s^{0.7}$; M2: $p = 1.20$; M3: $u_s u_\theta$ interaction.	Cross distance and incline over interior and endpoint settings, with mass varied to test the historical invariance.
`boyle_j_tube`	Boyle, trapped-air compression manuscript (2 October 1661)	$\mathcal{C}$: added pressure $P_{\mathrm{add}} \in [0, 220]$ kPa; thermal condition fixed to room temperature. $\mathcal{R}$: volume (mL), absolute pressure (kPa).	$V = 48\,\mathrm{mL}\,(P/101.3\,\mathrm{kPa})^{-1}$, with $P = 101.3\,\mathrm{kPa} + P_{\mathrm{add}}$.	Fixed amount of dilute gas; quasistatic isothermal equilibrium; calibrated end-space; no leakage, adsorption, or dissolution.	M1: $V \propto P^{-1} \to P^{-0.7}$; M2: $p = 0.80$; M3: bounded single-axis curvature $1 + 0.20u_P(1 - u_P)$.	Sweep added pressure from zero through two interior values to the upper endpoint after thermal equilibration.
`photoelectric`	Einstein (1905), Millikan's direct experiment (1916), and his Nobel lecture (delivered 1924)	$\mathcal{C}$: frequency $\nu \in [400, 1000]$ THz; absorbed photon flux $\phi \in [0.1, 2]$ relative. $\mathcal{R}$: stopping potential (V), saturation current (μA).	$V_{\mathrm{stop}} = \max(0, a\nu - \Phi/e)$, $a = 4.135667696 \times 10^{-15}$ V s, $\Phi/e = 2.28$ V; above threshold, $I_{\mathrm{sat}} = 3.0\phi$ μA.	Clean uniform cathode; monochromatic single-photon regime; complete collection; no contact drift, space charge, multiphoton response, or dark current.	M1: $h/e : 4.135667696 \times 10^{-15} \to 3.20 \times 10^{-15}$ V s; M2: $p = 1.20$; M3: $u_\nu u_\phi$ interaction.	Probe below and above the shifted threshold at widely separated frequencies and cross those settings with low and high flux.
`faraday_ induction`	Faraday, *Experimental Researches in Electricity*, vol. I (experiments 1831 onward; collected 1839)	$\mathcal{C}$: turns $N \in \{50, 100, 200, 400\}$; area $A \in [5, 100]$ cm^2; $\Delta B \in [1, 50]$ mT; ramp time $\Delta t \in [5, 500]$ ms. $\mathcal{R}$: peak emf (mV).	$\lvert\mathcal{E}\rvert = 20\,\mathrm{mV}(N/100)^{1} \times (A/20\,\mathrm{cm}^2)^{1} \times (\Delta B/10\,\mathrm{mT})^{1} \times (\Delta t/100\,\mathrm{ms})^{-1}$.	Uniform normal field; constant-slope ramp; identical turns; high-impedance readout; negligible loading, self-inductance, and stray flux.	M1: $\Delta t^{-1} \to \Delta t^{-0.7}$; M2: $p = 0.80$; M3: $u_N u_A$ interaction.	Sweep ramp duration across decades and cross turns with loop area while holding field change controlled.
`radioactive_ decay`	Rutherford and Soddy (1902, 1903)	$\mathcal{C}$: elapsed time $t \in [0, 300]$ min; active mass $m \in [0.25, 4]$ mg. $\mathcal{R}$: count rate (counts min^{-1}).	$R = 20 + 800(m/1\,\mathrm{mg})^{1} \times 2^{-t/(60\,\mathrm{min})}$ counts min^{-1}.	One radionuclide; common preparation and geometry; fixed background and efficiency; negligible self-attenuation and dead time.	M1: $m^1 \to m^{0.7}$; M2: $p = 1.20$; M3: $u_t u_m$ interaction.	Cross early/late times with low/high masses; retain a zero-time point to isolate mass scaling from decay.
`rutherford_ scattering`	Geiger–Marsden (1909), Rutherford (1911), and Geiger–Marsden (1913)	$\mathcal{C}$: angle $\theta \in [20, 160]^\circ$; $Z \in \{13, 29, 47, 79\}$; target density $n \in [0.1, 4]$ relative; exposure $\Phi \in [0.25, 2]$ relative. $\mathcal{R}$: count rate (counts min^{-1}).	$R = 120\,\mathrm{cpm}(Z/79)^{2} \times n^{1}\Phi^{1} \times [\sin(\theta/2)/\sin 30^\circ]^{-4}$.	Elastic nonrelativistic unscreened Coulomb scattering; thin single-scattering target; fixed beam energy and narrow detector acceptance.	M1: angular exponent $-4 \to -3$; M2: $p = 0.80$; M3: $u_\theta u_Z$ interaction.	Sweep angle broadly and cross it with at least two separated Z values; vary density/exposure to verify retained linear factors.

Continued on next page

Table 7 (continued)

Family	Historical anchor	Experimental interface	Historical relation	Validity scope	Counterfactual	Separating intervention
hooke_spring	Hooke, *Lectures de Potentia Restitutiva* (1678)	$\mathcal{C}$: force $F \in [0.5, 4]$ N; stiffness $k \in [20, 80]$ N m^{-1}. $\mathcal{R}$: extension x (cm).	$x = 5\,\mathrm{cm}(F/2\,\mathrm{N}) \times (k/40\,\mathrm{N\,m^{-1}})^{-1} \times [1 + c((F/2\,\mathrm{N})^2 - 1)]$, with $c = 0$.	Reversible quasistatic tensile regime; negligible hysteresis, creep, inertia, and temperature drift.	M1: cubic correction coefficient $0 \to 0.20$; M2: $p = 1.20$; M3: $u_F u_k$ interaction.	Sweep force at fixed stiffness to expose cubic curvature, then cross force and stiffness away from their reference values.
torricelli_efflux	Torricelli, *Opera geometrica* (1644)	$\mathcal{C}$: liquid head $h \in [0.1, 2]$ m; orifice area $A \in [1, 20]$ mm^2. $\mathcal{R}$: exit speed v (m s^{-1}).	$v = 4.43\,\mathrm{m\,s^{-1}} \times (h/1\,\mathrm{m})^{0.5} \times (A/5\,\mathrm{mm^2})^{0}$.	Steady incompressible lossless efflux; large reservoir; equal surface/outlet pressure; fixed opening shape and negligible contraction loss.	M1: $h^{0.5} \to h^1$; M2: $p = 0.80$; M3: $u_h u_A$ interaction.	Sweep head from low to high at fixed area, then vary area at matched heads to test both invariance and interaction.
ohm_circuit	Ohm, *Die galvanische Kette, mathematisch bearbeitet* (1827)	$\mathcal{C}$: voltage $V \in [0.2, 5]$ V; resistance $R \in [20, 500]\ \Omega$. $\mathcal{R}$: current I (mA).	$I = 10\,\mathrm{mA}(V/1\,\mathrm{V})^1 \times (R/100\,\Omega)^{-1}$.	Steady DC at fixed temperature; ideal source and meter; negligible self-heating, contact resistance, and parasitics.	M1: $R^{-1} \to R^{-0.5}$; M2: $p = 1.20$; M3: $u_V u_R$ interaction.	Sweep resistance over its full range at fixed voltage and cross low/high voltage with low/high resistance.
rc_discharge	Thomson's condenser-discharge analysis (1853); the benchmark uses its ideal zero-inductance RC limit	$\mathcal{C}$: $V_0 \in [2, 10]$ V; $R \in [25, 100]$ kΩ; $C \in [10, 40]$ μF; $t \in [0.05, 4]$ s. $\mathcal{R}$: capacitor voltage (V).	$V(t) = V_0 \exp[-a(t/RC)^q]$, with $a = 1$, $q = 1$, and $RC = R_{\mathrm{k}\Omega} C_{\mu\mathrm{F}}/1000$ s.	Ideal lumped first-order RC; constant components; high-impedance meter; negligible leakage, dielectric absorption, and parasitics.	M1: normalized-time shape $t^1 \to t^2$; M2: $p = 0.80$; M3: $u_{V_0} u_R$ interaction.	Sample several normalized times $t/(RC)$, then cross initial voltage and resistance while separately varying capacitance.
stefan_boltzmann	Stefan (1879) and Boltzmann (1884)	$\mathcal{C}$: emitter $T_e \in [400, 900]$ K; ambient $T_a \in [250, 350]$ K; area $A \in [20, 200]$ cm^2. $\mathcal{R}$: net radiative power (W).	$P = 35.439840119\,\mathrm{W} \times (A/100\,\mathrm{cm^2}) \times [(T_e/500\,\mathrm{K})^4 - (T_a/500\,\mathrm{K})^4]$.	Diffuse black emitter in a black enclosure with view factor one; uniform temperatures; calibrated exclusion of non-radiative transfer.	M1: linked $T_e^4, T_a^4 \to T_e^3, T_a^3$; M2: $p = 1.20$; M3: $u_{T_e} u_{T_a}$ interaction.	Sweep emitter temperature at fixed ambient, vary ambient at matched emitter temperature, and check area proportionality.
newton_cooling	Newton, *Scala graduum caloris* (1701)	$\mathcal{C}$: $T_0 \in [50, 100]^\circ$C; $T_a \in [10, 30]^\circ$C; $t \in [0, 25]$ min. $\mathcal{R}$: object temperature ($^\circ$C).	$T - T_a = (T_0 - T_a)e^{-kt}$, with $k = 0.12$ min^{-1} and fast-mode weight $w = 0$.	Lumped capacitance; constant ambient, transfer coefficient, and properties; negligible sensor lag and unmodelled loss channels.	M1: fast-mode weight $0 \to 0.25$; M2: $p = 0.80$ on ΔT; M3: $u_{T_0} u_{T_a}$ modulation of ΔT only.	Resolve the early-time curve at fixed endpoints, then cross initial and ambient temperatures while comparing temperature excess.
beer_lambert	Bouguer (1729), Lambert (1760), and Beer (1852)	$\mathcal{C}$: incident power $I_0 \in [20, 100]$ μW; concentration $c \in [0.05, 1]$ mM; path $l \in [0.25, 2]$ cm. $\mathcal{R}$: transmitted power (μW).	$I = I_0 e^{-\alpha c (l/1\,\mathrm{cm})^q}$, with $\alpha = 0.6$ mM^{-1} and $q = 1$.	Homogeneous stable non-scattering, non-fluorescent absorber in the linear regime; fixed wavelength, solvent, temperature, and calibrated windows.	M1: path exponent $l^1 \to l^2$; M2: $p = 1.20$; M3: $u_{I_0} u_c$ interaction.	Sweep path length at fixed concentration, then cross incident power and concentration at a shared path length.
michaelis_menten	Michaelis and Menten (1913); Haldane (1930)	$\mathcal{C}$: substrate $S \in [0.05, 10]$ mM; enzyme $E \in [0.25, 2]$ relative. $\mathcal{R}$: initial rate (μmol L^{-1} min^{-1}).	$v = V_{\max} ES/[K_m + S + \gamma S^2/K_m]$, with $V_{\max} = 80$ μmol L^{-1} min^{-1}, $K_m = 1$ mM, and $\gamma = 0$.	Single-substrate quasi-steady initial-rate window; negligible depletion, product, reverse reaction, inactivation, inhibition, and mass-transfer limitation.	M1: substrate-inhibition strength $0 \to 0.25$; M2: $p = 0.80$; M3: $u_S u_E$ interaction.	Sweep substrate through the rising, saturation, and high-substrate regimes; cross enzyme levels at shared substrate values.
coulomb_force	Coulomb, first memoir (presented 1785; cited from the 1789 collected volume)	$\mathcal{C}$: charges $q_1, q_2 \in [1, 10]$ nC; separation $r \in [2, 20]$ cm. $\mathcal{R}$: force magnitude (μN).	$F = 22.47\,\mu\mathrm{N} \times (q_1/5\,\mathrm{nC})^1 \times (q_2/5\,\mathrm{nC})^1 \times (r/10\,\mathrm{cm})^{-2}$.	Static point charges in vacuum; fixed localized distributions; no screening, polarization, image forces, or external fields.	M1: $r^{-2} \to r^{-1}$; M2: $p = 1.20$; M3: $u_{q_1} u_{q_2}$ interaction.	Sweep separation over its range, then independently cross both charges at a fixed separation.
kepler_orbit	Kepler, *Harmonices mundi libri V* (1619)	$\mathcal{C}$: semi-major axis $a \in [0.2, 8]$ AU; eccentricity $e \in [0, 0.7]$. $\mathcal{R}$: orbital period (days).	$P = 365.25\,\mathrm{days} \times (a/1\,\mathrm{AU})^{1.5}(1 + e)^0$.	Bound Newtonian two-body orbit about the same fixed central mass; negligible test mass, perturbations, tides, drag, and relativistic corrections.	M1: $a^{1.5} \to a^1$; M2: $p = 0.80$; M3: $u_a u_e$ interaction.	Sweep semi-major axis across more than one decade and vary eccentricity at a shared off-reference axis.

Continued on next page

Table 7 (continued)

Family	Historical anchor	Experimental interface	Historical relation	Validity scope	Counterfactual	Separating intervention
`brownian_diffusion`	Einstein, Brownian-motion paper (1905)	$\mathcal{C}$: lag $t \in [0.05, 5]$ s; tracer radius $r \in [0.2, 1]$ μm. $\mathcal{R}$: two-dimensional MSD (μm^2).	$\mathrm{MSD}_{2D} = 8\,\mu\mathrm{m}^2 \times (t/1\,\mathrm{s})^1 \times (r/0.5\,\mu\mathrm{m})^{-1}$.	Overdamped dilute spheres in a Newtonian fluid; fixed temperature and viscosity; far from walls; lag below drift/confinement times.	M1: $t^1 \to t^2$; M2: $p = 1.20$; M3: $u_t u_r$ interaction.	Sweep lag over two decades at fixed radius, then cross lag and radius at separated levels.
`poiseuille_flow`	Hagen (1839) and Poiseuille (preliminary reports 1840–1841; memoir issued 1844 and collected 1846)	$\mathcal{C}$: $\Delta P \in [2, 10]$ kPa; radius $r \in [0.25, 0.5]$ mm; length $L \in [20, 50]$ cm; viscosity $\mu \in [1, 3]$ mPa s. $\mathcal{R}$: flow rate (mL s^{-1}).	$Q = 1.2271846303\,\mathrm{mL\,s^{-1}} \times (\Delta P/10\,\mathrm{kPa})^1 \times (r/0.5\,\mathrm{mm})^4 \times (L/20\,\mathrm{cm})^{-1} \times (\mu/1\,\mathrm{mPa\,s})^{-1}$.	Steady fully developed incompressible Newtonian laminar flow; rigid straight circular tube; no-slip wall; losses measured only between pressure taps.	M1: $r^4 \to r^3$; M2: $p = 0.80$; M3: $u_{\Delta P} u_r$ interaction.	Sweep radius at fixed pressure, length, and viscosity; cross pressure and radius, then verify retained inverse length/viscosity factors.

Temporal sources. The current Temporal collection contains 24 families. Tables 8 gives interventions and sources. Collection batch and publication date are distinct quantities.

Table 8: **The 24 families in the Temporal collection**. These are derived from 23 primary papers and paired with benchmark-authored counterfactual responses. The source-reported branch is a deterministic implementation of the selected relation under the stated approximations. It does not reproduce all experiments or raw data in the paper. Numerical intervals and discrete sets below are the current legal query domains.

Family	Journal and source	Experimental interface	Source-reported relation	Validity scope	Counterfactual construction	Separating queries
Defect supersymmetric entropy	*Physical Review Letters* Huang et al. (2026b).	$\mathcal{C}$ inverse Rényi index $\gamma \in [0.5, 2]$. $\mathcal{R}$ normalized entropy y (dimensionless).	$y = 12[1 + 0.125(\gamma - 1)]$.	Symmetric $k = 2$ Wilson surface in the A_3 $(2, 0)$ theory. $r_1 = r_2 = 1/2$. Exact relation at fixed theory and representation.	M1 replaces γ by γ^2, preserving $y(1) = 12$.	Query $\gamma = 0.5$ and 2, and retain $\gamma = 1$ as a common reference.
Turbulent scaling exponent	*Physical Review Letters* Mailybaev & Thalabard (2026).	$\mathcal{C}$ moment order $p \in [1.4, 4.6]$. $\mathcal{R}$ scaling exponent ζ_p (dimensionless).	$\zeta_p = p/3 - Cp(p-2)\epsilon^2$, $C = (2^{2/3} + 1)/(4\,2^{1/3} \ln 2)$ and $\epsilon = 0.1$.	Inertial-range shell-model asymptotics through $O(\epsilon^2)$. Terms $O(\epsilon^4)$ are omitted. Continuous moment orders are benchmark queries.	M1 changes ϵ from 0.1 to 0.2. The correction is multiplied by four and $\zeta_2 = 2/3$ is retained.	Sample moment orders below and above $p = 2$, including 1.4 and 4.6. The protected $p = 2$ alone does not separate the laws.
MnF_2 crossover displacement	*Physical Review Letters* Ohlendorf et al. (2026).	$\mathcal{C}$ strain $\epsilon_{xy} \in \{-0.0004, -0.0008, -0.0011, -0.0015\}$ and field $B \in [0.644, 5.796]$ T. $\mathcal{R}$ crossover shift ΔT^* (K).	$T_c = 67.467 - 59\epsilon_{xy} - 0.01585B^2$. $\Delta T^* = T_c[0.56(0.67171381472)(-B\epsilon_{xy})^{2/3}/T_c]^{3/2}$.	Compression along [110], field along c. Mean-field exponent $\beta = 1/2$. All admitted strains are negative, so $-B\epsilon_{xy} = \lvert B\epsilon_{xy}\rvert$.	M1 replaces every B by $W_{1-\epsilon_{xy}/0.0015}(B)$, including its occurrences in T_c.	Cross strain levels with interior fields, including $B = 3.22$ T. Endpoint fields are preserved by the composition.
Confined ciliate swimming	*Physical Review Letters* Antunes et al. (2026).	$\mathcal{C}$ mean slip $u \in [-0.4, 0.4]$ and amplitude $a \in [0.24, 1.36]$. $\mathcal{R}$ normalized laboratory-frame mean swimming velocity $\langle v_p \rangle / v_m$. All dimensionless.	$\langle v_p \rangle / v_m = \sqrt{\max[(1-u)^2 - a^2, 0]} - 1$.	Resonant channel and metachronal wavelengths, lubrication and incompressible Stokes flow. Fixed reduced mobility $\lvert G_n \rvert = 2$. Includes the phase-locked plateau.	M1 and M2 replace u and a by $W_2(u)$ and $W_2(a)$.	Cross interior slip and amplitude settings through the oscillatory-to-locked transition. Include interior values because corners alone are preserved.
Quantum interconnect rate	*Physical Review Letters* Luo et al. (2026).	$\mathcal{C}$ modes $M \in \{2, 4, 8, 11\}$, storage efficiency $\eta_s \in [0.54, 0.86]$ and noise $p_n \in \{0.01, 0.10, 0.50\}$. $\mathcal{R}$ scaled rate $\widetilde{R} = (L_0/c)R_{\mathrm{qm}}$ (bits).	$A = 1 - (M-1)p_n/M$. $\mathcal{I} = \log_2 M + A \log_2 A + (M-1)(p_n/M)\log_2(p_n/M)$. $\widetilde{R} = 0.7e^{-250/44}\mathcal{I}\eta_s^4/[27(2-\eta_s)(4-3\eta_s)]$.	Ideal 1000-km, four-segment repeater with two nesting levels. $N = \eta_d = 1$, $L_{\mathrm{att}} = 22$ km. Equal input probabilities and depolarizing memory noise. Lifetime losses omitted.	M1 replaces all η_s occurrences by $W_{1+p_n/0.1}(\eta_s)$, retaining the same information term.	Cross noise and mode levels at interior $\eta_s = 0.7$. Efficiency endpoints alone are preserved.

Continued on next page

Table 8 (continued)

Family	Journal and source	Experimental interface	Source-reported relation	Validity scope	Counterfactual construction	Separating queries
Time-dependent CP asymmetry	*Physical Review Letters* LHCb Collaboration (2026).	$\mathcal{C}$ proper time $t \in [1.25, 7.25]$ ps. $\mathcal{R}$ integrated CP asymmetry A_{CP} (dimensionless).	$A_{CP}(t) = 0.82\sin(0.507t) + 0.13\cos(0.507t)$.	Helicity-angle-integrated $B^0 \to K^0_S\mu^+\mu^-$ asymmetry. $\Delta\Gamma_d = 0$ approximation and fixed dilepton-mass exclusions. Proper time is an observation coordinate.	M1 replaces both time occurrences by $W_2(t)$.	Sample interior proper times across the oscillation, in addition to 1.25 and 7.25 ps. The two endpoints alone coincide.
Fermentation minimum selling price	*Nature Communications* Bhagwat et al. (2026).	$\mathcal{C}$ yield $Y \in [0.28, 0.92]$ g product/g sugar and titer $T \in [21, 109]$ g product/L. $\mathcal{R}$ minimum selling price P (2019 USD/kg product).	$P = [0.311 + 0.353/Y + 10.9/T + 0.356/(YT)]/0.983$.	Fitted techno-economic relation for glucose-to-succinic acid, low-pH index 25. Productivity 0.66 g L^{-1} h^{-1} and recovery 0.983. Large-scale fed-batch plant model.	M1 and M2 replace yield Y and titer T by $W_2(Y)$ and $W_2(T)$ in every occurrence.	Cross interior yield and titer levels with their endpoints to distinguish both reciprocal terms and their interaction.
Shear-thinning cutting force	*Nature Communications* Zhao et al. (2026).	$\mathcal{C}$ $r \in \{0.6856, 0.9775, 1.0829, 1.4455, 1.5566\}$ μm. $v \in \{0.083, 0.125, 0.167, 0.208, 0.250, 0.292, 0.333, 0.375, 0.417, 0.458, 0.500\}$ m/s. $h \in \{5, 23, 33\}$ mm. $\mathcal{R}$ relative force $y = F/F_{\mathrm{ref}}$ (dimensionless).	$y = (r/0.6856)^{0.968}(v/0.5)^{0.72}(h/33)^{0.344}$.	Printed empirical fit for the fixed optimized starch fluid and calibrated straight blades. Discrete joint query grid. Other compositions, blade geometries and the unlabelled 16-mm-sheet series are excluded.	M1 replaces the radius exponent 0.968 by $0.968(v/0.5)$, coupling radius and speed. The common normalization is retained.	Cross at least two nonreference radii with low and high speeds, with depth fixed. The reference radius or reference speed alone hides the mutation.
Off-diagonal scattering density	*Physical Review Letters* Köhnes et al. (2026).	$\mathcal{C}$ cross section $\sigma \in [0.04, 0.36]$. $\mathcal{R}$ local probability-density value p_σ. Both coordinates are dimensionless.	$m = 0.967^2/(2(0.967) + 7.013)$, $z = \sigma/m$, $a = (2(0.967) - 1)/(2(0.967) + 7.013)$. $p_\sigma = e^{-z}[1 + a((4 - z)z - 2)]/m$.	Leading and first subleading Heidelberg asymptotic density for $\beta = 2$, $T_1 = T_2 = 0.967$ and absorption 7.013. Used near the Ericson transition. The finite query interval is not a separately normalized support.	M1 composes p_σ with $W_2(\sigma)$. M2 multiplies it by $W_2'(\sigma) = 2(\sigma - 0.04)/0.32$.	Query endpoints and interior cross sections. The Jacobian changes density heights even where the coordinate composition fixes an endpoint.
Thin-film differential reflectance	*Physical Review Letters* Puri et al. (2026).	$\mathcal{C}$ SnS thickness $d \in [72.05, 252.45]$ nm. $\mathcal{R}$ AC-polarized differential reflectance $\Delta R/R$ (dimensionless).	$R = \lvert(r_{01} + r_{12}z)/(1 + r_{01}r_{12}z)\rvert^2$, $z = e^{4\pi i\widetilde{n}d/800}$, $r_{01} = (1 - \widetilde{n})/(1 + \widetilde{n})$, $r_{12} = (\widetilde{n} - 1.7276)/(\widetilde{n} + 1.7276)$. $\Delta R/R = -0.00302\,\partial_n \ln R$ at $\widetilde{n} = n + 0.12i$, $n = 4.23$.	Normal-incidence 800-nm light in air/SnS/MgO. First-order real-index perturbation only, with imaginary-index change omitted. All multiple reflections and the analytic derivative are retained.	M1 replaces every thickness occurrence by $W_2(d)$ in the complete differentiated Fresnel response.	Sweep interior thickness through interference extrema. The original thickness endpoints are common to both laws.
Squeezed-vacuum energy surface	*Physical Review Letters* Ramírez et al. (2026).	$\mathcal{C}$ bosonic coordinates $q, p \in [-1.6, 1.6]$. $\mathcal{R}$ thermodynamic-limit energy h_∞ in the fixed model energy unit.	$h_\infty = (q^2 + p^2)/2 - 2\sqrt{1 + 0.04q^2(q^2 + p^2)}$.	Mean-field squeezed-vacuum surface after eliminating atomic coordinates. $\omega = 1$, $\omega_0 = 2$, $\gamma/\omega = 0.4$, below spectral collapse. Not an exact finite-system energy.	M1 and M2 replace all q and p occurrences by $W_2(q)$ and $W_2(p)$.	Cross interior positive, zero and negative phase-space coordinates to test curvature and the source reflection symmetries.
Pulsed magnetic-field sensing	*Physical Review X* Tabatabaei et al. (2026).	$\mathcal{C}$ field $B \in [10^{-8}, 9 \times 10^{-8}]$ T and pulse blocks $N \in \{10, 30, 58, 80, 100\}$. $\mathcal{R}$ signal-to-noise ratio.	$t_c = 1.78 \times 10^{-6} N$ s. $A_N = \sqrt{60/(t_c + 0.004)} \times e^{-t_c/0.000104}/\mathcal{N}$. $\mathrm{SNR} = A_N \sin(0.34\gamma_e B t_c)$, $\gamma_e = 176086268505.04355$ s^{-1} T^{-1} and $\mathcal{N} \simeq 14.109773394162719$.	Fixed acquisition and noise parameters of the nonlinear sensitivity model. Sinusoidal field at a filter-function lobe and sensing duration much shorter than T_1. $\mathcal{N}$ is evaluated from the frozen source noise expression.	M1 replaces B only inside the phase by $W_{1+N/100}(B)$. Acquisition, decay and noise terms retain their forms.	Cross interior field strengths with distinct block counts. Field endpoints alone conceal the composition.

Continued on next page

Table 8 (continued)

Family	Journal and source	Experimental interface	Source-reported relation	Validity scope	Counterfactual construction	Separating queries
Undetected-light visibility	*Nature Communications* Severs Millard et al. (2026). Advance publication.	$\mathcal{C}$ object-arm intensity transmission $T \in [0.1, 0.9]$. $\mathcal{R}$ fringe visibility V. Both dimensionless.	$a = 0.784^2(0.47)T$. $b = 0.8(0.118)$ and $c = 0.885^2 + 0.8$. $V = \frac{2\sqrt{0.8a}(0.885+\sqrt{b})}{b+c(1+a)}$.	Michelson nonlinear thin-film model with undepleted pump and seed. Stimulated emission dominates. Only the first four-wave-mixing reflection and within-window interference are retained.	M1 replaces both transmission occurrences by $W_2(T)$.	Scan interior transmission values between 0.1 and 0.9. The two endpoints alone agree.
Davemaoite phase boundary	*Nature Communications* Hao et al..	$\mathcal{C}$ pressure $P \in [36, 124]$ GPa. $\mathcal{R}$ transition temperature T_{tr} (K).	$T_{\mathrm{tr}} = 533 + 2.4P$.	Printed fitted tetragonal-to-cubic phase boundary for lower-mantle davemaoite. Deterministic use of central fit coefficients. Fit uncertainties are not per-query noise.	M1 replaces P by $W_2(P)$.	Compare at least two interior pressures against the source straight line, with endpoints as shared references.
CoO orbital magnetoresistance	*Science* Schmitt et al. (2026).	$\mathcal{C}$ remaining Cu* thickness $t \in [3.88, 5.32]$ nm. $\mathcal{R}$ orbital-magnetoresistance amplitude A (percent).	$A = 0.32 + 0t$ percent.	Approximate plateau of CoO(5 nm)/Cu* at 150 K. One Hall bar is progressively ion-milled and exposed to air for 12 h after each etch. Thickness is resistance-calibrated.	M1 changes the affine slope from 0 to 0.08 percent/nm, giving $A = 0.32 + 0.08t$.	Compare low and high remaining thickness to test the plateau against a nonzero slope under the fixed protocol.
MnF_2 transition baseline	*Physical Review Letters* Ohlendorf et al. (2026).	$\mathcal{C}$ loading-path strain $\epsilon \in \{-0.0015, -0.0011, -0.0008, -0.0004\}$ and field $B \in \{0, 6.44\}$ T. $\mathcal{R}$ baseline T_c (K).	$T_c = 67.467 - 59\epsilon - 0.01585B^2$.	Fitted non-altermagnetic baseline along compressive [110] loading, including co-induced symmetric strain. Not independently imposed pure shear or the shifted crossover T^*.	M1 replaces B^2 by B^4/B_0^2, with fixed $B_0 =$ 1 T. The field coefficient retains units K/T^2.	Compare $B = 0$ and 6.44 T at matched strain, then repeat across strain levels.
Quasiperiodic localization exponent	*Physical Review Letters* Lin et al. (2026a).	$\mathcal{C}$ comprises couplings $\lambda_1, \lambda_2 \in [0.20, 0.67]$. $\mathcal{R}$ is the Lyapunov exponent L. All quantities are dimensionless.	$\Lambda = \frac{\lambda_2(1+\sqrt{1-\lambda_1^2})}{\lambda_1(1+\sqrt{1-\lambda_2^2})}$. $L = \max(0, \ln \Lambda)$.	The source-theory unitary almost-Mathieu relation is evaluated at non-Hermiticity $\eta = 0$, irrational quasiperiodic field $\Phi = (\sqrt{5}-1)/2$ and phase $\theta = 0$. Positive L gives inverse localization length. A zero value covers extended or critical states. The interface returns the asymptotic theoretical exponent without finite-time photon-counting fluctuations.	M1 and M2 replace every occurrence of λ_1 and λ_2 by $W_2(\lambda_1)$ and $W_2(\lambda_2)$.	Cross distinct interior couplings with $\lambda_2 > \lambda_1$, including $(0.25, 0.50)$. Include the reversed pair as a zero-response control. Equal couplings and domain corners alone do not separate the branches.
Viscous–inertial suspension rheology	*Physical Review Letters* Konidena et al. (2026).	$\mathcal{C}$ is the Stokes number $\mathrm{St} \in [0.01, 300]$. $\mathcal{R}$ is normalized macroscopic friction $r = \mu/\mu_c$. Both are dimensionless.	$r = 1 + 14.3\sqrt{\frac{0.065(1+0.029\,\mathrm{St})}{1+0.1\,\mathrm{St}}}$.	The source-reported dense-suspension relation is restricted to interparticle friction $\mu_f = 0.15$ and the constructed packing-related slice $K_\phi = J + 0.1I^2 = 0.065$, with $\mathrm{St} = I^2/J$. The displayed coefficients specify the fixed deterministic response. Changing Stokes number does not independently scan packing or particle friction, and the admitted interval is not extrapolated.	M1 replaces both Stokes-number occurrences by $W_2(\mathrm{St})$.	Sweep logarithmically separated interior values such as 0.1, 10 and 100. Compare the crossover curvature and use the shared endpoints only as controls.

Continued on next page

Table 8 (continued)

Family	Journal and source	Experimental interface	Source-reported relation	Validity scope	Counterfactual construction	Separating queries
Splashing-drop rim response	*Physical Review Letters* Riboux & Gordillo (2026).	$\mathcal{C}$ is $\mathrm{We} \in [288, 592]$. $\mathcal{R}$ comprises rim speed $v_t = V_t/V$ and rim diameter $h_t = H_t/R_d$. All quantities are dimensionless.	$\mathrm{Re} = \sqrt{\mathrm{We}}/0.0105$, $A = \sqrt{3}/(2\mathrm{Re})$, $b = 1/\mathrm{We}$. Let $x > 0$ solve $x^4 - bx - A = 0$ and set $t_e = x^2$. $v_t = 0.45\sqrt{3/t_e}$, $h_t = 2.2t_e^{3/2}$.	The calibrated low-viscosity, high-Reynolds-number approximation describes the early two-jets spray regime. The fixed layer depth is $H/R_d = 1.8$, Ohnesorge number 0.0105, and gas density and viscosity ratios 0.001 and 0.018. Both readouts describe the toroidal rim at $t_0 = 0.07$ in units of R_d/V, including ballistic rim motion and early thickening.	M1 replaces every Weber-number occurrence by $W_2(\mathrm{We})$ in the linked root construction and both outputs.	Sample several interior Weber numbers and compare the coupled speed and diameter curves. Preserve the low and high endpoints as common references rather than using them alone.
Trapped-ion optical Magnus response	*Physical Review Letters* Leindecker et al. (2026).	$\mathcal{C}$ comprises focal-plane positions $x, y \in [-1.04, 1.04]$ μm and transition category $q \in \{-1, +1\}$. $\mathcal{R}$ is ground-state survival probability P_q, dimensionless.	$g_q = (x/w)(1 + qy/z_R)e^{-(x^2+y^2)/w^2}$, $w = 1.3\,\mu\mathrm{m}$, $z_R = \pi w^2/(0.729\,\mu\mathrm{m})$. $a = \sqrt{1 + 2(w/z_R)^2}$, $g_* = \frac{a+1}{2\sqrt{2e}}e^{-(w/z_R)^2/(a+1)^2}$. $P_q = \cos^2[\pi\lvert g_q\rvert/(2g_*)]$.	A resonant single pulse addresses a fixed $^{40}\mathrm{Ca}^+$ ion initially in $m_g = +1/2$. The beam propagates along z, is polarized along y, and the quantization field points along $-x$. The first correction beyond the paraxial quadrupole-gradient model is retained. Each transition uses its modeled peak coupling over the one-waist square for its own fixed π-pulse calibration.	M1 replaces both occurrences of x by $W_2(x)$. The coordinate y, discrete transition category and fixed pulse calibration are retained.	Cross interior horizontal positions, especially $x = 0$, with positive and negative vertical positions and both transition categories. Track the survival-profile displacement while retaining the horizontal endpoints as controls.
Parity-prepared oscillator squeezing	*npj Quantum Information* Lin et al. (2026b). Advance publication.	$\mathcal{C}$ comprises coherent-displacement amplitude $a \in [0.14, 1.26]$ and per-step intensity transmissivity $\eta \in [0.865, 0.985]$, both dimensionless. $\mathcal{R}$ is final conditional squeezing S in dB.	$\Pi_\beta = D(\beta)[1 + (-1)^{\hat{n}}]D^\dagger(\beta)/2$, $\mathcal{M}_\beta(\rho) = \mathcal{L}_\eta(\Pi_\beta\rho\Pi_\beta)$ with full pure loss $\mathcal{L}_\eta$. Normalize $\rho_* \propto \mathcal{M}_0\mathcal{M}_{-ia}\mathcal{M}_{ia}(\lvert 0\rangle\langle 0\rvert)$. $S = -10\log_{10}[4\min_\theta \mathrm{Var}_{\rho_*}(x_\theta)]$, $x_\theta = (\hat{a}e^{-i\theta} + \hat{a}^\dagger e^{i\theta})/2$. The card gives the exact nine-exponential reduction.	Three ideal displaced-parity measurements start from vacuum and retain only the all-even outcomes. Full unobserved pure loss follows each measurement, including the third, and the final state is normalized by its full success probability. The response retains coherences and all loss events. The displacement amplitude is not time, and no Gaussian-state or single-loss approximation is used.	M1 replaces all 19 amplitude occurrences by $a^2/1.4$. This uses the preregistered source amplitude reference, not W_2 on the smaller public interval.	Cross low, middle and high admitted amplitudes with several transmissivities. Compare the complete squeezing curve. The public amplitude endpoints are not fixed by this composition.

Continued on next page

Table 8 (continued)

Family	Journal and source	Experimental interface	Source-reported relation	Validity scope	Counterfactual construction	Separating queries
Mean-field microcanonical coexistence	*Physical Review X* Serbyn et al. (2026).	$\mathcal{C}$ comprises $h_x \in [0.06, 0.14]$ and $h_z \in [-0.24, -0.16]$. $\mathcal{R}$ comprises crossing energy density $\epsilon_\times$, common magnetization magnitude $m_\times$ and signed branches m_z^-, m_z^+. All are dimensionless in the stated Pauli and J_3 units.	$s = -\sqrt{(-h_z + \sqrt{3}h_x)/3}$. $\epsilon_\times = 2s^3$, $m_\times = \frac{2}{3}\sqrt{-3h_z + \sqrt{3}h_x}$. $m_z^\pm = \frac{s \pm \sqrt{s^2 - 8h_z/3}}{2}$.	The uniform thermodynamic LMG-3 mean-field problem maximizes microcanonical entropy at fixed energy $\epsilon = z^3 + h_z z + h_x x$ over $x^2 + z^2 \le 1$. The selected negative-energy crossing has two global entropy maxima with opposite signed longitudinal magnetizations and the same magnitude. The Pauli magnetization convention is used. This target excludes finite-size disorder, tunneling rates and dynamical ETH hybridization thresholds.	M1 replaces every h_x occurrence in all four outputs by $0.05 + (h_x - 0.05)^2/0.10$. The fixed references are the original safety-box endpoints 0.05 and 0.15.	Cross transverse and longitudinal fields over the legal box. Compare the crossing energy and both signed branches together. The current public transverse-field endpoints do not remain invariant.
Loss-mediated transfer and entanglement	*Physical Review A* Ren & Li (2026).	$\mathcal{C}$ is scaled time $\tau = \kappa t \in [30, 270]$. $\mathcal{R}$ comprises $P_{L\|L}$, $P_{R\|L}$, C_L, $P_{L\|R}$, $P_{R\|R}$, C_R, all dimensionless. The suffix specifies the independent initial qubit excitation, and C_d is reduced two-qubit concurrency.	$u = 2^{-1/4}/10$, $\mathbf{a}^{(d)} = e^{-iH_{\mathrm{eff}}\tau}\mathbf{e}_d$. $H_{\mathrm{eff}} = \begin{pmatrix} 0 & 0 & u \\ 0 & 0 & u \\ u & u & -(1+i)/2 \\ u & -iu & 0 \end{pmatrix}$ in basis (L, R, c_1, c_2). $P_{j\|d} = \|a_j^{(d)}\|^2$, $C_d = 2\|a_L^{(d)} a_R^{(d)}\|$. Amplitudes are unnormalized and lost probability occupies the vacuum.	The engineered sideband-selected rotating-wave model retains both equal-loss cavities with detunings $\Delta_{1,2}/\kappa = \pm 1/2$ and relative coupling phase $-\pi/2$. Each reset starts with one specified qubit excited and both cavities in vacuum. Cavity recycling is unconditional. Intrinsic qubit decoherence and thermal excitation are omitted. No cavity elimination or exactly vanishing reverse transfer is imposed.	M1 replaces all 48 time occurrences in the six complete modal outputs by $W_2(\tau)$. The populations and concurrence require no derivative factor.	Sample several interior times from both preparations, including early and late transients. Compare transferred populations and concurrence jointly. The interval endpoints remain shared.
Non-Markovian cooperative radiation	*Quantum* González & Rivas (2026).	$\mathcal{C}$ is scaled time $\tau = \gamma_0 t \in [2, 18]$. $\mathcal{R}$ comprises unconditional Dicke populations P_2, P_1, P_0 and signed atomic energy-loss rate $j = I/(\omega_0\gamma_0)$. All four are dimensionless.	$\partial_\tau \rho = -i[H, \rho] + \mathcal{D}[b]\rho$, $H = (J_- b^\dagger + J_+ b)/\sqrt{2}$, $\rho(0) = \|ee, 0\rangle\langle ee, 0\|$. $\mathcal{D}[b]\rho = b\rho b^\dagger - \{b^\dagger b, \rho\}/2$. $P_n = \langle n\| \mathrm{Tr}_b\, \rho \|n\rangle$ for symmetric Dicke states $n = 0, 1, 2$. $j = -\partial_\tau(2P_2 + P_1)$. The card retains every modal coefficient.	Two identical atoms have common coupling phase, a vacuum Lorentzian reservoir, linewidth $\lambda/\gamma_0 = 1/2$ and transition ratio $\omega_0/\gamma_0 = 1000$. The resonant rotating-wave and narrowband pseudomode model retains all reachable excitation sectors and loss recycling. The initial atomic state is fully excited. Negative signed intensity represents reabsorption, and the three populations sum to one.	M1 composes all populations and signed intensity with $g(\tau) = W_2(\tau)$. M2 multiplies the composed intensity by $g'(\tau) = (\tau - 2)/8$, preserving its derivative definition.	Sample interior early and late times and compare populations with signed energy flow. Endpoint populations agree between branches, whereas endpoint intensities need not because of the derivative factor.

D LLM DISCLOSURE

Generative AI tools were used solely to assist in summarizing existing literature and refining the manuscript for grammar and readability. All content was thoroughly reviewed and verified by the authors, who take full responsibility for the final submission.