# CAUSAL DISCOVERY BENCH: DYNAMIC SANDBOX FOR LLM CAUSAL DISCOVERY AND INFERENCE EVALUATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

As large language model (LLM) agents get actively adopted for autonomous research, they must not only reason over existing evidence, but also design informative experiments, allocate resources, analyze sequential observations, and decide when to stop and make conclusions. Yet these capabilities remain poorly measured. We introduce Causal-Discovery-Bench, a verifiable sandbox with configurable structural equation models and resource constraints to evaluate sequential experimental design and causal discovery without domain-semantic cues. Across 3,322 episodes with 13 models from 5 vendors, we analyze agent behavior along 40 metrics in seven families. Agents' performance exhibits a four-level difficulty hierarchy spanning single-experiment design, sequential exploration of causal relationships, resource-aware experimentation balancing cost and statistical power, and autonomous resolution of conflicting evidence. We identify a substantial performance gap between causal inference from correctly curated evidence and navigating causal discovery through autonomous experimentation: Analyzing agents' action trajectories attributes 93.9% of their mistakes to experimental design rather than to inference, which is further consolidated with ablation studies. Relaxing resource constraints alone does not reliably resolve failures in sequential discovery. In contrast, basic statistical tools substantially improve performance across multiple model families, while advanced tools provide little additional benefit when models lack prerequisite lower-level capabilities.

**Outline**

## 1 INTRODUCTION

The fast growing capability of Large Language Model (LLM) agents has motivated strong interest in deploying them for autonomous scientific research, machine learning engineering agents, and recursive self-improvement (RSI). Such use cases require an LLM agent to actively navigate through research questions under limited budgets, turning observations into hypotheses, hypotheses into experiments, and experimental results into both revised conclusions and new directions of exploration. Generating evidence via autonomous sequential exploration, rather than merely reasoning over supplied evidence, becomes an increasingly sought-after capability.

*Can an LLM agent self-navigate the design and analysis of a sequence of experiments using potentially limited resources to solve open questions?*

Existing research reveals failures in interactive discovery and long-horizon autonomous research but does not diagnose agents' missing capabilities and failure modes when handling such tasks under finite data-acquisition budgets. We built CAUSAL-DISCOVERY-BENCH, a verifiable sandbox to shed light on why LLM agents struggle and to design experiments that elicit how they can be strengthened without training. Running with configurable structural equation models and resource constraints, CAUSAL-DISCOVERY-BENCH provides interactive environments for LLM agents to perform data-driven causal discovery tasks through sequential experiments and automatically evaluates their task performance against the ground truth. Variable names CAUSAL-DISCOVERY-BENCH



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

are deliberately randomized to rule out memory effects from domain-semantic cues, so that task performances reflect agents' true causal discovery.

We find that the agents struggle to design sequential experiments to collect evidence rather than to interpret existing evidence, and that this single deficit persists through every intervention that does not free agents from experiment design. Provided with adequate controlled experiment data, a mid-tier production model identifies the correct causal network structure in 14 of 15 sandboxes; however, when the same volume of data is collected but with inadequate variable control measuring total rather than direct treatment, the success number drops to 9. Across all 3,322 scored episodes, 93.9% of attributed errors occur because agents draw conclusions before collecting adequate evidence to support final conclusions. Increasing the resource budget by $25x$ increases success rates in identifying causal network structure from 17% to 78%. However, the chances for the agents to perform proper controlled experiments remain at 0.38 per episode, showing that accuracy is achieved by blindly exploring more candidates rather than by properly isolating their treatment effects.

Delegating part of the experiment design to the interface is more effective than prompt guidance. Averaged across prompt conditions, including a statistical-power-calibrated sample size cap increases the frequency of successful controlled experiments from $0.75$ to $8.22$ per episode; averaged across sample size caps, the strongest prompt intervention increases the experiment control success rate only from $4.78$ to $5.82$. In a separate study, a five-step procedure induces controlled experimentation in three models that never exhibit it unprompted, although it does not close their final task performance gap with stronger models.

## 2 Related Work

**Autonomous research and long-horizon agents.** LLM agents have been increasingly deployed to plan and execute scientific experiments (Boiko et al., 2023). However, evaluations of autonomous research and professional workflows reveal persistent difficulties in checking intermediate results, pivoting from unsuccessful approaches, and maintaining task objectives (Fei et al., 2026; Zhu et al., 2026). Agents can even underperform simple test-time scaling methods based on random retries (Wang et al., 2026). Interpreting these results is challenging because objective measures of research quality are often unavailable. Moreover, variation in agents' domain knowledge can complicate standardized comparisons across models. We address these limitations with an interactive sandbox that provides known causal ground truth, removes domain-specific semantic cues, and supports controlled ablations of budgets, tools, and instructions.

**Interactive model discovery and causal reasoning benchmarks.** Several interactive benchmarks are relevant but differ in the evidence agents receive, the resources they may spend, and the task itself: BoxingGym evaluates predictive model discovery in domain-grounded environments (Gandhi et al., 2025), CausalGame reasoning under hidden confounding, selection bias and measurement error (Chen et al., 2026), and IGDA edge queries answered with binary feedback (Havrilla et al., 2025). CausaLab is closest, since it also uses hidden SCMs and budgeted interventions, but evaluates broad graph and equation recovery alongside held-out prediction (Yang et al., 2026). Our emphasis is diagnostic and prescriptive. We isolate where an agent's experimental design, statistical power and causal inference break down, and we test what repairs them. We build on CausalProfiler's SCM infrastructure for explicit assumptions, known ground truth, and sequential interaction with a persistent population (Panayiotou et al., 2025).

## 3 Causal Discovery as Sequential Experimentation

We evaluate LLM agents' ability to design and analyze a sequence of experiments for causal discovery. Each task takes place in a sandbox that simulates an online A/B testing environment, where a population of instances produces a noisy outcome $Y$ influenced by a subset of variables $\mathbf{X} = (X_1, \ldots, X_p)$. Agents must use experimental evidence to identify which variables have a direct causal effect on $Y$ and estimate the magnitudes.

The causal structure is represented by a directed acyclic graph (DAG) $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, with vertices $\mathbf{V} = V_1, \ldots, V_{p+1}$ corresponding to $V_j = X_j$ for $j \leq p$ and $V_{p+1} = Y$. A vertex $V_i$ is an *ancestor*

of $V_j$ if a directed path leads from $V_i$ to $V_j$, and a *parent* of $V_j$ if $(V_i, V_j) \in \mathbf{E}$. Each directed edge represents a direct causal effect. The discovery objective is therefore to identify the parents of $Y$.

Each sandbox contains a finite population of $N$ instances governed by a linear structural equation model (SEM) with instance-specific coefficients. Let $\mathrm{Pa}(j)$ denote the index set of the parents of $V_j$. For instance $n$, the structural equation is

$$V_j^n := \sum_{i \in \mathrm{Pa}(j)} \big(a_{ij} + \delta_{ij}^n\big) V_i^n + \varepsilon_j^n, \tag{SCM}$$

where $a_{ij} \neq 0$ is the baseline coefficient shared across the population, $\delta_{ij}^n$ is an exogenous coefficient deviation, and $\varepsilon_j^n$ is additive exogenous noise. All coefficient deviations and additive noise terms are mutually independent across edges, vertices, and instances. For each edge $(i, j)$, the deviations $\{\delta_{ij}^n\}_{n=1}^N$ are independently drawn from an identical zero mean distribution; likewise, for each vertex $j$, the noise terms $\{\varepsilon_j^n\}_{n=1}^N$ are independently drawn from an identical zero mean distribution across instances. For a vertex with no parents, the equation reduces to $V_j^n = \varepsilon_j^n$.

At initialization, each instance independently receives a realization of the additive noise vector $\boldsymbol{\varepsilon}^n = (\varepsilon_1^n, \ldots, \varepsilon_{p+1}^n)$ and the coefficient deviations $\boldsymbol{\delta}^n := \{\delta_{ij}^n : (V_i, V_j) \in \mathbf{E}\}$. Both remain fixed throughout the task and are reused whenever the instance is sampled for an observation or experiment. Together with the baseline coefficients and Equation equation SCM, they determine the instance's observational values and its outcomes under intervention.

Agents interact with the sandbox through a sequence of *observations* and *experiments*, subject to a fixed total budget $B$. Both actions incur costs linear in sample size: an observation of $m$ instances costs $c_o \cdot m$, while an experiment involving $m$ instances across both intervention groups costs $c_e \cdot m$, where $c_e > c_o > 0$ are the respective per-instance costs. For each action, the agent selects a sample size $m$ and samples $m$ instances uniformly with replacement from the population. Let $n_1, \ldots, n_m$ denote their indices, allowing repetitions. An observation returns the sampled data $\big((X_1^{n_j}, \ldots, X_p^{n_j}, Y^{n_j})\big)_{j=1}^m$. For an experiment, the agent additionally selects a variable index set $\mathcal{S} \subseteq 1, \ldots, p$ and two intervention vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, each of dimension $|\mathcal{S}|$. The sampled instances are randomly divided into two equally sized groups to respectively receive the interventions $\mathrm{do}((X_i)_{i \in \mathcal{S}} = \mathbf{x}_1)$ and $\mathrm{do}((X_i) i \in \mathcal{S} = \mathbf{x}_2)$. Each intervention replaces the structural equations for the selected variables with their assigned values. The remaining variables and outcome are then computed using Equation SCM and each sampled instance's fixed noise realization $\boldsymbol{\epsilon}^n$ and $\boldsymbol{\delta}^n$. Computed results are shared to the agent.

The agents' task is to identify the direct parents of $Y$ and estimate the direct causal effect of each parent. With $V_{p+1} = Y$, the target parent set is

$$\mathcal{P}_Y := \{j \in \{1, \ldots, p\} : (V_j, V_{p+1}) \in \mathbf{E}\}.$$

For each $j \in \mathcal{P}_Y$, the target effect is the average change in $Y$ induced by a unit increase in $X_j$ while holding the remaining variables $\mathbf{X} - j$ fixed:

$$\theta_j(x; \mathbf{z}) := \mathbb{E}[Y \mid \mathrm{do}(X_j = x + 1, \mathbf{X}_{-j} = \mathbf{z})] - \mathbb{E}[Y \mid \mathrm{do}(X_j = x, \mathbf{X}_{-j} = \mathbf{z})]\,.$$

Here, the expectation is over the data-generating distribution of $\boldsymbol{\varepsilon}^n$ and $\boldsymbol{\delta}^n$ for $n = 1, \cdots, N$. Under the linear SEM in Equation equation SCM and the assumption $\mathbb{E}[\delta_{ij}^n] = 0$, this controlled direct effect becomes independent of $x$ and $\mathbf{z}$ and equals the baseline coefficient, i.e.

$$\theta_j(x; \mathbf{z}) \equiv \tau_j = a_{j,p+1}, \tag{1}$$

In our task, the agent must return an estimated parent set $\widehat{\mathcal{P}}_Y$ and an effect estimate $\widehat{\theta}_j$ for each selected parent, using a sequence of observations and experiments whose cumulative cost does not exceed the budget $B$. This is highly challenging due to the following considerations.

### 3.1 Challenges in identifying direct parents and their treatment effects.

To identify whether a variable $X_j$ is indeed a parent of $Y$, agents must design controlled experiments or select observational adjustment methods that isolate each candidate variable's direct effect. Without prior knowledge of the parent set or the intervening causal structure (Pearl, 2009; Cinelli et al., 2024), this task is quite challenging, as the agent has to resolve several potential confusions.

**Correlation vs Causality** The agent needs to distinguish between spurious correlation and genuine causality, which requires combining observations with experiments properly. For example, as long as a common ancestor is shared, $X_j$ and $Y$ can demonstrate a correlation without any causal relationship. In such cases, the agent needs to perform proper controlled experiments to rule out $X_j$ from $Y$'s ancestors.

**Parents vs Ancestors** The agent also needs to distinguish the parents of $Y$ from the ancestors. Changing $X_j$ from $x'$ to $x$ while leaving the other variables to follow their structural equations yields the total average treatment effect

$$\tau_j(x, x') := \mathbb{E}\big[Y \mid \mathrm{do}(X_j = x)\big] - \mathbb{E}\big[Y \mid \mathrm{do}(X_j = x')\big] = (x - x') \sum_{\pi:, X_j \rightsquigarrow Y} \prod_{(u,v)\in\pi} a_{uv}{}^{1}. \quad (2)$$

The total treatment effect can be misleading for identifying direct parents: it is often nonzero as long as $X_j$ is an ancestor $Y$.

**Direct vs Total Treatment Effects from Confounders** Even if the agent has correctly identified $X_j$ as a direct parent, it remains nontrivial to calculate the direct treatment effect if $X_j$ is a confounder shared with $Y$ and some other direct parent(s) of $Y$. In rare cases, such a confounder can have a zero total treatment effect on $Y$ if its direct treatment effect cancels out with its effect channeled through other direct parents.

In addition to designing a sequence of observations and experiments that can, in principle, resolve these puzzles, agents must understand that they work in an environment where heterogeneous coefficients and noisy outcomes can introduce sampling uncertainty. They must quantify uncertainty in their effect estimates and choose sample sizes that provide sufficient statistical power for their inference within the available budget.

### 3.2 Challenges from finite action budgets

The finite budget makes causal discovery a sequential experimental design problem. Each action must be chosen using the accumulated evidence and the remaining resources. Agents face four interrelated decisions: which causal hypotheses to test next; how to design observation or experiment sets that identify direct effects; how many samples to allocate to each test given its cost and statistical power; and how to update causal conclusions and resolve conflicting evidence. These decisions require balancing the breadth of hypothesis testing against the precision of individual effect estimates, while determining whether further data collection is warranted before submitting a final conclusion.

## 4 The Sandbox Design

The sandbox implements the sequential causal discovery task defined in Section 3. Its design separates population generation, agent interaction, and evaluation: the simulator generates observations and experimental outcomes, the agent chooses how to collect and interpret evidence, and the evaluator compares the resulting submission and trajectory against the true causal model.

### 4.1 Sandbox generation

We develop a diverse set of sandboxes to stress test the agents in distinctive ways to draw insights on their task performance traits and failure modes. Among these sandboxes, we consider five special families of DAGs with special topological structures: chain, star, fork, collider, and diamond (Figure 2), to test whether the agent can handle particular reasoning and planning tasks. We also enrich our sandbox diversity with another family of DAGs generated by a random graph sampling procedure. We further consider different DAG sizes in each family to vary task difficulty.

In our sandboxes, we don't model hidden causality links, i.e., all variables modeled in the SEMs can be observed by the agents. However, only a subset of the variables can be intervened with the `do` operator, with the subset being configurable for each sandbox. We also label variable names with

[1] The $\{\delta_{ij}^n\}_{i,j,n}$ are zeroed by the expectations as they are mutually independent and each has zero mean.

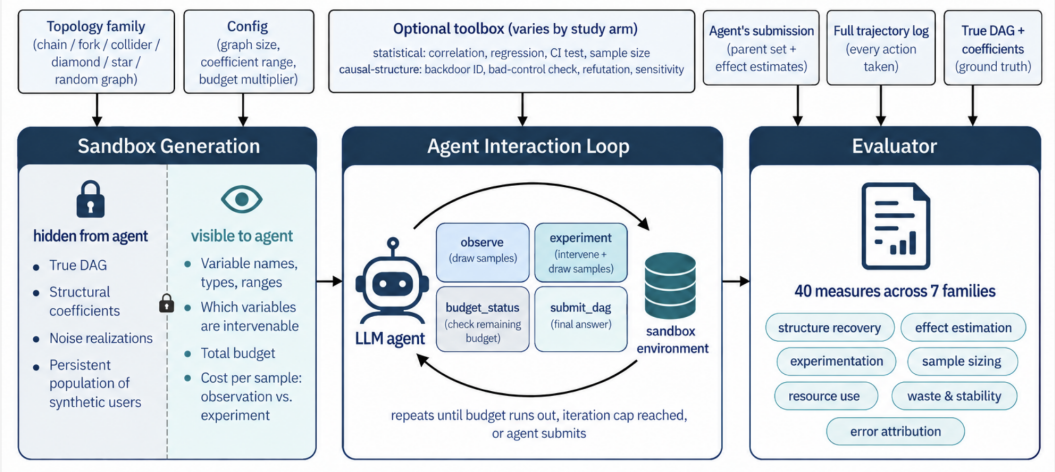


Figure 1: The sandbox pipeline. Left: generation splits every instance into a hidden half (the true causal model) and a visible half (the interface the agent actually sees). Middle: the agent interacts through four priced core actions in a loop, optionally equipped with a statistical or causal-structure toolbox depending on the study arm, never seeing the hidden half directly. Right: the evaluator scores the final submission and the full trajectory against the hidden ground truth, producing forty measures across seven families.

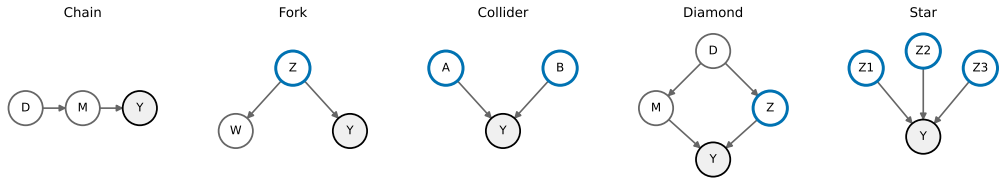


Figure 2: Canonical examples for our five DAG families. Blue circles are parents of the outcome $Y$

randomly sampled strings that carry no domain semantic meaning so that our sandboxes measure the agents' raw causal reasoning powers without spillovers from memorized domain semantic cues that can be given away linguistically from variable names (Zečević et al., 2023).

Given the DAG structure, each sandbox extends CausalProfiler (Panayiotou et al., 2025) to initialize its parameters in closed-form SEMs and its persistent population. These parameters provide the evaluator with ground truth used to score agents' performances. We provide details on this initialization procedure in Appendix D.

We built another version to draw observation and experiment samples with replacement for benchmark robustness check. Study results between both samplers differ by less than 4%. (Appendix H).

### 4.2 Agent Interface

The sandbox sends its task to an agent using a prompt in which the agent sees the data description that a machine-learning engineer would typically get when starting a project: The names and pointers to variables $\mathbf{X}$ and outcome $Y$, as well as their data type, range, and whether they can be intervened upon with the `do` operation. The total budget for observation and experimentation actions and per sample costs are also provided. The DAG, structural coefficients, and noise realizations remain hidden from the agent. The agent works on the task through multiple turns of interactions with the sandbox, where at any turn it can take any of the four actions: `observe` to see samples, `experiment` to intervene with samples and see raw results, `budget_status` to check the re-

maining budget, and `submit_dag` to conclude the task by submitting the parents of $Y$ and their per-unit treatment effects to the evaluator.

We log agents' full task trajectory, including their actions and outputs at each turn, and feed the trajectory to the evaluator.

### 4.3 Evaluator and Metrics

The evaluator assesses an agent's task performance using a comprehensive metric system that covers different dimensions of an agent's ability to carry out a causal discovery task. The metric system examines the agent's entire task trajectories in a given set of sandboxes and evaluates not only the accuracy of final answers but also the progressive performances at each turn, including statistical power for sample sizes, budget efficiency, variable exploration coverage, etc. These dimensions are quantified by forty metrics that can be grouped into seven metric families (further details in Appendix G).

The seventh family is what turns a wrong answer into a missing capability. It charges every wrong edge back to the evidence the agent had gathered for that candidate, and those states form four levels that the task requires in order: (i) designing one experiment that identifies the effect, (ii) covering every candidate, (iii) sizing each test to the power the budget can buy, and (iv) resolving the evidence once it conflicts. The levels are ordered rather than merely distinct, because failing at an earlier one hides the later ones: an agent that never tests a candidate produces no evidence that could be misread, so its error can only be charged to the level at which it stopped. An agent's profile over the four is therefore read from the bottom up, and a level is only informative once the ones below it are passed.

Our metric system also evaluates task difficulty based on budget scarcity by considering the ratio between available budgets and the heuristic budget $B_{\text{test}}$ needed to complete the task using a strong algorithm, which consists of a screen followed by one adequately powered controlled contrast per candidate without using the true graph (details in Appendix B). Particularly, we set a default budget allocation of

$$B = 5B_{\text{test}} \qquad \text{(Default Budget)}$$

for our sandbox and treat this budget as a baseline in later experiments.

## 5 Experiments and Findings

### 5.1 Where can an agent fail

We run a set of experiments to understand why an agent can conclude with wrong parents or estimate treatment effects with poor accuracy. We focus on how `Claude Sonnet 4.6` and apply the budget rule at Default Budget throughout the experiments. We summarize the analysis perspectives provided by these experiments in Table 1.

| **Ablation study** | **Detailed setup** |
|---|---|
| `Estimation-only` | Estimating treatment effects given DAG; `15 sandboxes` |
| `Inference-only` | Identifying parents and estimating treatment effects given a set of observation and experiment data; `30 sandboxes`, `15 control + 15 treatment` |
| `DAG Topology` | Full causal discovery task on different DAG families; `318 sandboxes` |
| `DAG Sizes` | Full causal discovery task with different graph sizes (6-18 vertices) in each DAG family; `70 sandboxes` |

Table 1: Ablation studies used to localize agent failures.

In the `estimation-only` experiment, we ablate the DAG-structure discovery component from our task by directly showing the DAG in each sandbox. Controlled interventions that separately vary each parent are sufficient for identification. However, `Claude Sonnet 4.6` did not conduct any controlled experiments in any of the 15 task trajectories we ran with it. This shows the agents' fundamental limitation in conducting controlled experiments.

In the `Inference-only` experiment, we ablate sequential experimentation entirely and provide a fixed evidence package comprising one observation and one experiment dataset for each candidate parent variable, although the agent is still allowed to make additional observations and experiments. We evaluate two matched arms across the same 15 sandboxes. In the controlled arm, each candidate is subjected to a `do` operation with all other candidates held fixed, allowing the experiment to identify direct effects. In the treatment arm, the focal candidate is still applied with a `do` operation, but the remaining variables are uncontrolled, so the experiment measures total effects. The agent recovers the exact parent set in 14 of 15 controlled-evidence trajectories but only 9 of 15 treatment-evidence trajectories.

These two studies show that agents can usually reach the right conclusion when they are given the right evidence. Once supplied with misleading (though true) evidence, the correct identification of parents drops from 14 to 9, signaling that the agent struggles to make independent adjustments regarding evidence quality. If required to design sequential experiments, the accuracy of treatment effect estimation drops for correctly identified parents, with the mean absolute error increasing from $0.11$ to $0.32$. This result explains why 94.5% of the errors in our broader evaluation over 3,779 agent task trajectory episodes occur before the final inference step. If an agent never collects useful evidence, it has no information to interpret incorrectly.

In the `DAG sizes` experiment, we keep the original causal discovery task fixed while increasing the size of random graphs from 6 to 18 variables. The agent fails to recover the exact parent set at every size, but the source of difficulty changes systematically. The mean number of unresolved candidates increases from $4.6$ to $8.5$, whereas recall among candidates the agent resolves remains approximately constant ($0.48$ at six variables and $0.44$ at eighteen, with no monotonic trend between them). Increasing the graph size, therefore, primarily reduces how much of the problem the agent investigates, rather than how accurately it judges the variables it encounters.

In the `DAG topology` experiment, we vary the graph across five canonical structures. Across 2,842 trajectories, uncontrolled contrasts account for only $1\%$ of attributed errors in graphs without mediation ($n = 1{,}738$), but $44\%$ in graphs with mediation ($n = 1{,}104$). Never-tested errors move in the opposite direction, decreasing from $69\%$ to $42\%$. Graph topology, therefore, determines which capability the task demands: mediation tests whether the agent uses appropriate controls, whereas graphs without mediation primarily expose failures in hypothesis coverage.

These two experiments separate two sources of difficulty. Graph size determines how much of the task the agent leaves unfinished, while graph topology determines which causal-reasoning skill is required to finish it.

### 5.2 How models differ

Running thirteen models in the same set of sandboxes with the same prompt, we find both common and distinctive failure modes. Models demonstrate distinctive failure modes for more advanced experimental designs: one model makes errors mostly due to neglecting candidates from experimentation, while another keeps running experiments without controlling variables.

**Neither model tier nor vendor strongly inform causal discovery capability.** A flash model reaches 93% exact recovery and finishes above every pro-tier model in the panel, whereas another pro model reaches 73% (Figure 3, right). This ordering persists as we increase sandbox task difficulty, where the same flash model scores 91% against 36% for a pro-preview model of its own family. Two vendors appear at both ends of the ordering, so neither the parameter tier nor the model family predicts task performance.

**Controlled experimentation distinguishes task performance.** Two models conduct about ten controlled experiments per episode, whereas the other four conduct at most one; no model falls between these extremes (Figure 3, center). Among the 40 measures in Appendix G, only this behavior effectively divides the models into strong vs weak groups. Accuracy, recall, and budget use, instead, vary continuously. Final accuracy also obscures how answers were reached: the two weakest models resolve only 2% and 5% of candidate relationships, yet achieve exact parent-set accuracies of 13% and 33%, respectively. They can therefore reach correct answers without collecting the identifying evidence needed to support them.

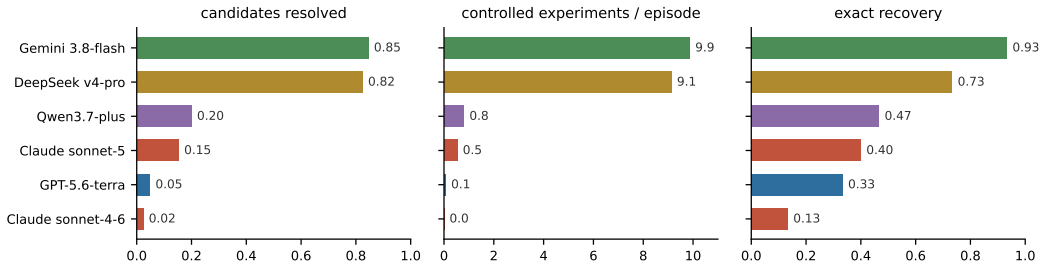


Figure 3: Budget improves coverage without inducing controlled experimentation. **Left:** Exact parent-set recovery across budgets and toolboxes (see Section 5.3 for budget and tool ablation). **Center:** Number of controlled experiments conducted under the same conditions. **Right:** Distribution of attributed errors in the no-tools condition. As the budget increases, errors from untested candidates decline, while errors from experiments without appropriate controls become dominant.

### 5.3 How to Fix Agents

We tried four ways of improving the agent without changing the model: more budget, analysis tools, a tighter interface, and a written procedure. The two that supply resources leave its behavior where it was; the two that supply the planning step change it, and still leave it short of a model that plans for itself.




Figure 4: Left: exact recovery against budget for each toolbox, on the same instances. Four statistical primitives at $\times 1$ already match no tools at $\times 125$, while the causal-structure toolbox tracks no tools throughout. Right: where the remaining errors sit for the no-tools arm, as a percentage of attributed errors. Budget does not remove the failure; it moves it from never-tested to tested-without-control. Vermilion marks the budget lever and the coverage failure it removes; blues mark what has to install the controlled contrast rather than pay for it, throughout.

**Budget and tools buy more testing, not better testing.** A budget 125 times larger lifts exact recovery from 17% to 78%, while the number of controlled experiments never passes 0.30 per episode, against about nine for a model that runs them by habit (Figure 4). What changes is which mistake the agent makes: errors from never testing a candidate fall from 90% of the total to 0%, and errors from testing one without holding the others fixed rise from 5% to 78%. A statistical toolbox reaches the same accuracy for a fraction of that budget by reading cheap observational rows rather than buying interventions, and leaves the same contrast undone. A causal-structure toolbox changes nothing at either end of the range: it checks conclusions, and the agent has not produced one to check.

**Capping the purchase size makes the agent test properly.** Capping how much a single call may buy raises exact recovery from 21% to 64% and controlled experiments from 0.75 to 8.22 per episode, while rewording the prompt moves the same measures only from 35% to 48% and from 4.78 to 5.82; under the tight cap the three prompts become indistinguishable (Figure 5). The cap

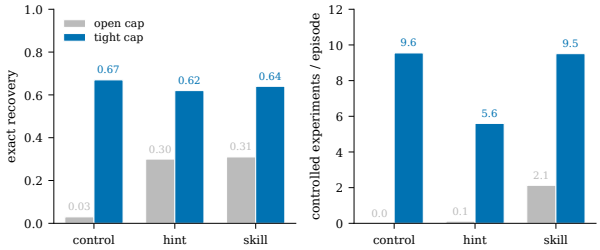


Figure 5: Cap against prompt, on the instances of the preregistered $3 \times 2$. Under the open cap the three prompts are far apart; under a cap set to the instance's power requirement they are indistinguishable and every one of them runs controlled experiments. The environment, not the wording, decides whether the behavior appears.

works by choosing the sample size the agent will not choose for itself, since recovery peaks at a cap of 7 against a power requirement of 6.58. A written five-step procedure has the same effect without touching the interface, taking three models from no controlled experiments at all to between 2.7 and 4.5 per episode. Their accuracy still lands at 87%, 73% and 40%, short of the 100% reached by the model that runs controlled experiments without being told to. The behavior is installable; the capability is not.

Budget buys coverage: more candidates tested, none of them settled. The only levers that buy the contrast itself are the two that take a decision away from the agent, the cap by fixing the sample size and the procedure by stating the comparison, while the toolbox that checks conclusions never meets one. And a contrast the environment forces is not a contrast the model plans: the three scaffolded models stay thirteen to twenty-seven points behind.

## 6 Conclusion

We develop Causal-Discovery-Bench, a sandbox in which an agent is required to find the direct causes of an outcome by buying observations and interventions against a finite budget. Every action it takes is recorded and scored on forty measures, and every cause it gets wrong is traced to the step it failed at: designing an experiment that identifies the effect, covering every candidate, sizing each test to the power it needs, or resolving evidence that conflicts. Across 3,322 episodes from thirteen models, almost every failure happens at one of the first three steps and not at the fourth. Agents resolve the evidence they hold and fail to produce evidence worth holding.

Our analysis identifies the following failure patterns and potential fixes. Causal inference needs a controlled comparison, varying one candidate while holding the rest fixed. More budget and better tools get the agent to test more candidates, and it still never holds the others fixed while it tests one. Two things do make it hold them fixed: a cap on how much one purchase may buy, and a written procedure in the prompt. The cap supplies the sample size the agent does not compute, and the procedure states the comparison it does not construct. Neither, however, makes the agent as accurate as a model that needs no such help, and that gap in accuracy is visible only in what the agent did to reach its answer.

Rather than only simply evaluating the final answer an agent submits, our method reads a signal off the trajectory by charging for every action and recording which of the four steps an error came from. That record tells an agent whose evidence supports the causes it names from one whose evidence does not. It is also a denser training target than the answer, since it marks which action in the sequence was the wrong one instead of only that the result was wrong. Nothing in it depends on causal graphs, so the same scoring applies past direct-cause discovery to any task where evidence has to be bought before it can be read. As agents are handed more research that requires collecting new evidence, that is the sandbox worth building, one that prices every action and reads the trajectory against the four capabilities the task demands in order.

## AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

We used generative AI tools (Claude, Anthropic; Codex, OpenAI) for tasks that require disclosure: implementing and refining our experiment infrastructure of Appendix H; processing unstructured experiment data to draw insights, such as in Section 5.3; writing offline analysis scripts that recompute reported metrics from stored trajectories; and helping to interpret results, including several explanatory paragraphs in Section 5.3. We did not use them to generate synthetic data, to build theoretical or conceptual frameworks, to formulate mathematical claims, or to write proofs: the data-generating process, the reference cost in Appendix B, and the derivations in Appendix A are the authors' own. Translation, dataset cleaning , and thematic analysis do not apply to this work.

Under recommended disclosure, we also used them to draft and edit parts of Sections 4.1 and 5, convert text to LaTeX, refer to prior work. We checked every reported statistic against the raw episode records rather than accepting AI output as given, and verified AI-suggested citations against the sources themselves. We take responsibility for the final content of this work.

## REFERENCES

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

Zhenhao Chen, Yongqiang Chen, Chenxi Liu, Junchi Yu, Xiangchen Song, Zijian Li, Jialin Li, Philip Torr, Bo Han, and Kun Zhang. CausalGame: Benchmarking causal thinking of LLM agents in games, 2026. URL https://arxiv.org/abs/2607.04293.

Carlos Cinelli, Andrew Forney, and Judea Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, 53(3):1071–1104, 2024.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15:3741–3782, 2014.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, 2000.

Yanlin Fei, Nazhou Liu, Xinmiao Yu, Shaolong Chen, Lei Li, Rahul Thapa, Madalina Ciobanu, Navan Preet Singh, Qingqing Mao, and Ritankar Das. How do agents fail on AutoResearch: End-to-end diagnostic evaluation on 100 real-world frontier research tasks, 2026. URL https://arxiv.org/abs/2608.14905.

Kanishk Gandhi, Michael Y. Li, Lyle Goodyear, Louise Li, Aditi Bhaskar, Mohammed Zaman, and Noah D. Goodman. BoxingGym: Benchmarking progress in automated experimental design and model discovery, 2025. URL https://arxiv.org/abs/2501.01540.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *PMLR*, 2018.

Alex Havrilla, David Alvarez-Melis, and Nicolo Fusi. IGDA: Interactive graph discovery through large language model agents, 2025.

Murat Kocaoglu, Alexandros G. Dimakis, and Sriram Vishwanath. Cost-optimal learning of causal graphs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PMLR*, 2017.

Panayiotis Panayiotou, Audrey Poinsot, Alessandro Leite, Nicolas Chesneau, Marc Schoenauer, and Özgür Şimşek. CausalProfiler: Generating synthetic benchmarks for rigorous and transparent evaluation of causal machine learning, 2025. URL https://arxiv.org/abs/2511.22842.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

Yike Wang, Huaisheng Zhu, Zhengyu Hu, Yige Yuan, Zhengyu Chen, Shakti Senthil, Hannaneh Hajishirzi, Yulia Tsvetkov, Pradeep Dasigi, and Teng Xiao. Rethinking the evaluation of harness evolution for agents. In *COLM 2026 The 2nd Workshop on Lifelong Agents: Learning, Aligning, and Evolving*, 2026.

B. L. Welch. The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35, 1947.

Junlin Yang, Dylan Zhang, Xiangchen Song, Qirun Dai, Xiao Liu, Yuen Chen, Aniket Vashishtha, Jing Shi, Chenhao Tan, and Hao Peng. CausaLab: A scalable environment for interactive causal discovery toward AI scientists, 2026. URL https://arxiv.org/abs/2605.26029.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023.

Liya Zhu, Jingzhe Ding, Jian Zhang, Jianbo Xue, Shihao Liang, Ge Zhang, Yi Zhu, Duju Zeng, Xiang Gao, Qingshui Gu, Mailun Gao, Huimin Che, Yan Zhao, Peiheng Zhou, Haojun Wang, Chaobo Xian, Lili Le, Chi Wu, Yiwei Liu, Shengda Long, Jiale Yang, Fangzhi Xu, Sijin Wu, Haodong Duan, Chao He, Zhaojian Li, Minchao Wang, Huan Zhou, Jiani Hou, Chuqian Yu, Weiran Shi, Hongwan Gao, Jiamin Chen, Guanhong Chen, Tingqin Luo, Kaiyuan Zhang, Zhixin Yao, Qing Hua, Yuhao Jiang, Jin Chen, Pu Chen, Zhenyu Hu, Xingyu Li, Zhengxuan Jiang, Meng Cao, Tianfeng Long, Haozhe Wang, Mingzhang Wang, Yichen Zhang, Yiming Dai, Chenchen Zhang, Jiaying Wang, Xinying Liu, Xingzu Liu, Lingling Zhang, Xinjie Chen, Yujia Qin, Wangchunshu Zhou, Zhiyong Wu, Yang Liu, Jiaheng Liu, Lei Zhang, Shen Yan, Wenhao Huang, Zaiyuan Wang, and Xiaolong Chang. Workflow-GYM: Towards long-horizon evaluation of computer-use agentic tasks in real-world professional fields, 2026. URL https://arxiv.org/abs/2606.11042.

## A Deferred derivations

**Every ancestor has a non-zero total effect.** Two directed paths from $X_j$ to $Y$ use distinct edge sets, so no monomial in the sum of equation 2 cancels another and the sum is a non-zero polynomial in $\{a_{uv}\}$ whenever at least one directed path exists. The zero set of a non-zero polynomial has Lebesgue measure zero, and the generator draws each $a_{uv}$ from a continuous range, so the event that equation 2 vanishes for an ancestor of $Y$ has probability zero. This is the cancellation that faithfulness rules out by assumption; here the sampling scheme rules it out instead.

**The two routes, and the cost of one conclusive contrast.** Pinning $X_{-j} = \mathbf{X} \setminus \{X_j\}$ to zero removes every path but the edge itself, so that

$$\Delta_j \;:=\; \tfrac{1}{2}\Big[\mathbb{E}\big(Y \mid do(X_j{=}{+}1,\, X_{-j}{=}0)\big) - \mathbb{E}\big(Y \mid do(X_j{=}{-}1,\, X_{-j}{=}0)\big)\Big] \;=\; a_{jY}, \qquad (3)$$

and $\Delta_j = 0$ iff $X_j \notin \mathrm{Pa}(Y)$; only a variable that is both a descendant of $X_j$ and an ancestor of $Y$ has to be neutralised, since anything else cancels across the arms. Because the generator places no variable outside $\mathbf{X} \cup \{Y\}$, the same quantity is also identified by conditioning: a set of observed variables can block every back-door path, and regressing $Y$ on $X_j$ given that set returns $a_{jY}$ from observational rows at $c_{\mathrm{obs}}$ each. Where a variable that must be neutralised cannot be manipulated, only the second route is open, which is why we also report recall restricted to intervenable parents.

Pinning leaves the residual variance of $Y$ in equation 3 exactly $\sigma^2 = \mathrm{Var}(\varepsilon_Y)$. Sizing a two-arm contrast to detect a difference of $\delta_{\min}$ at $80\%$ power and $\alpha = 0.05$ gives $n^* = 16\sigma^2/\delta^2_{\min}$ rows per arm (Cohen, 1988), where $\delta_{\min}$ is the smallest coefficient the generator can draw. The two arms sit at $\pm 1$, so the difference actually to be detected is $2\delta_{\min}$ and this allocation is deliberately conservative. A conclusive test therefore costs $2n^* c_{\exp}$ against total budget.

**Why the observational route is not available.** A linear model with non-Gaussian noise is identifiable from $P(V)$ alone by independent-component methods (Shimizu et al., 2006), and the exogenous noise here is uniform. The route is nonetheless closed: each row carries its own per-edge coefficient deviation (Section 4.1), so the population is not a single fixed linear model and the identifiability result does not apply to it. Even where it did, recovering $\mathbf{G}$ would leave every $a_{jY}$ unestimated, and the rows it reads are charged against total budget like any others.

## B THE BUDGET FORMULA

**The formula.** Collecting the two phases gives the reference cost the body refers to,

$$B_{\text{test}} := 10(p+1)c_o + \min(2k, p)\,\frac{16\sigma^2_{\text{ref}}}{\delta^2_{\min}}\,c_e, \tag{4}$$

whose first term buys the observational screen over $p$ candidates and the outcome, and whose second buys one adequately powered controlled contrast for each of $\min(2k, p)$ of them. Every instance is priced by equation 4 and given $B = 5B_{\text{test}}$ by default.

**Derivation.** Prior work on budgeted experiment design fixes a budget and asks what structure is learnable inside it (Kocaoglu et al., 2017; Ghassami et al., 2018); we invert the question and price the strategy a competent agent who does not already know the graph would have to run. It proceeds in two phases. It first screens: on $p$ variables it buys $10p$ observational rows and runs a constraint-based search over them (Colombo & Maathuis, 2014). It then tests each surviving candidate with one two-arm intervention powered at $80\%$, which by Cohen's formula needs $n^* = 16\sigma^2/\delta^2_{\min}$ rows (Cohen, 1988), where $\delta_{\min}$ is the smallest coefficient the generator can draw and $\sigma^2$ is the noise on the edge under test, the variance left once the other candidates are held fixed, taken at its largest over the instance. Because a constraint-based screen on finite samples returns false neighbors, we let it hand back up to $\min(2k, p-1)$ candidates.

**What the formula does and does not price.** It prices the *controlled* route, because $\sigma^2$ is measured with the other candidates held fixed; an agent that intervenes without controlling also carries the variance the other parents contribute, so the same power costs it more. It prices one *test* rather than one instance, because $80\%$ power per candidate is not $80\%$ power per instance: with several candidates to settle, a multiplicity correction stands between the two. Across the benchmark the median $B_{\text{test}}$ is 31.7 (IQR 23.8–48.5) and one powered contrast takes a median of 6.6 rows.

**The reference statistician.** The procedure is: screen, one controlled contrast per candidate, Welch $t$-test (Welch, 1947), Bonferroni over the candidate tests. Unlike the pricing formula it uses no knowledge of $k$ and tests every candidate, so it may run more tests than $B_{\text{test}}$ pays for. Run on 15 instances with 20 sampling repetitions each, it recovers the exact parent set in 299 of 300 trials at the default $5B_{\text{test}}$ and in 60 of 300 at $B_{\text{test}}$, which is what calibrates the default: a competent strategy succeeds there and a fifth of it fails. Because the procedure is conservative in the sense just described, both figures understate what the budget affords.

**The variance proxy.** $\sigma^2_{\text{ref}}$ in equation 4 is the largest empirical variance of the per-unit coefficient deviations on an edge into $Y$, plus the variance of the optional additional outcome-noise term. It normalizes the budget across instances; it is not the sampling variance of any particular experimental design, and equation 4 is correspondingly a reference price rather than a sufficient allowance.

## C MODELS

Every model is the vendor's hosted endpoint, called with default sampling settings and no system prompt beyond the task prompt of Section 4.2. Names are as the vendor published them at the time of the run; endpoints for several have since been retired, so the model string in the released episode records is authoritative.

| Model | Model string | Studies |
|---|---|---|
| Claude Sonnet 4.6 | `claude-sonnet-4-6` | single-model studies, panels |
| Claude Sonnet 4.6, thinking | `claude-sonnet-4-6-thinking` | thinking ablation |
| Claude Sonnet 5 | `claude-sonnet-5` | workhorse panel |
| Claude Opus 5 | `claude-opus-5` | frontier panel |
| Claude Fable 5.1 | `claude-fable-5-1` | frontier panel |
| GPT-5.6 Terra | `gpt-5.6-terra` | scaffolding cross, panels |
| GPT-5.6 Sol | `gpt-5.6-sol` | frontier panel |
| GPT-6 Astra | `gpt-6-astra` | scaffolding cross, panels |
| Gemini 3.8 Flash | `gemini-3.8-flash` | scaffolding cross, panels |
| Gemini 3.1 Pro Preview | `gemini-3.1-pro-preview` | frontier panel |
| DeepSeek V4 Pro | `deepseek-v4-pro` | scaffolding cross, panels |
| Qwen3.7 Plus | `qwen3.7-plus` | scaffolding cross, panels |
| Qwen3.8 Max | `qwen3.8-max` | frontier panel |

## D INSTANCE CONSTRUCTION

**The answer key and the realized population.** The evaluator reads the parent set from the DAG and each direct effect from the baseline SEM coefficient, so the key is exact by construction. The average effect realized in a finite population can still differ from that target, because the per-unit coefficient deviations of equation SCM need not average to zero over $N$ units; we draw a large population in each sandbox to keep the discrepancy below the reporting precision of every effect measure.

**Free instances.** $p$ is drawn uniformly from a difficulty-tier range and edges are placed by a fixed-order Erdős–Rényi rule: node $i$ receives $\mathrm{Binomial}(i-1, \pi)$ parents sampled without replacement from $X_1, \ldots, X_{i-1}$, with $\pi$ set so the expected edge count matches a target that scales with $p$, from $\log p$ at the sparsest tier to $p(p-1)/4$ at the densest. The fixed topological order rules out cycles by construction. One endogenous node is then designated $Y$, and the whole graph is re-drawn until that node's in-degree and number of upstream colliders both fall inside the study's target range. These are the two structural knobs Section 5 varies independently.

**Canonical instances.** The five shapes of Figure 2 are placed as an explicit edge list parameterized by one integer level: fork, for example, is $D \to Y$ plus one $Z_i \to Y$, $Z_i \to W_i$ pair per level, and similarly for the other four. Sweeping one shape's size while every other identification descriptor stays fixed is what a free random graph cannot guarantee on demand.

**Mechanism family.** The engine also supports non-linear and tabular mechanisms. Every study reported here uses the linear family, which is what keeps the effects closed-form and supplies the $\sigma^2$ that the budget rule of Section 4.3 depends on.

## E RUNNER AND TERMINATION

**Provider abstraction.** Every provider is driven through one interface: the runner emits a provider-agnostic list of messages and tool specifications and receives back a normalized response carrying reasoning text, at most one tool call, a stop reason and token usage. Provider-specific payload that must round-trip verbatim to keep a multi-turn tool-call chain valid, such as reasoning content some providers require replayed on the next turn or a thought signature others attach to a tool call, is carried as opaque data the runner never inspects. Responses carrying several tool calls in one turn are normalized to at most one. This uniform contract, backed by two adapter implementations, is what lets one harness run every model in Section 5 without a per-model branch in the loop.

**Termination.** An episode ends one of four ways: the agent submits; the budget reaches zero; the iteration cap is reached; or the adapter fails. Exhausting the budget or the iteration cap does not end the episode outright. The agent is given one further turn on which observation and experiment are unavailable but submission is still possible. An agent that never submits is scored as having submitted the empty set rather than excluded from the corpus.

## F INSTANCE DESCRIPTORS

Twenty-five descriptors are recorded for every instance, grouped by the obstacle they belong to. *Identification* (seven): topology, confounders, colliders, mediation chains, longest chain, non-intervenable variables, intervenable-parent fraction. *Estimation* (five): smallest true effect, largest distractor effect, effect spread, signal confusion, $n^*$. *Design cost* (four): $B_{\text{test}}$, two sufficiency ratios, the budget multiplier. A further five record the scale of the search and four the provenance of the instance, so that any instance in the corpus can be regenerated exactly.

## G THE FORTY MEASURES

All forty are written by one code path for every fully scored episode. Nine cover structure recovery, of which *exact recovery* is reported throughout; two cover effect estimation; six attribute each wrong edge to the evidentiary state the agent had reached for that candidate — never tested, tested without control, underpowered, unresolved at submission, or resolved and misread; eleven describe what the agent ran, of which *controlled experiments per episode* is reported throughout; two cover sample sizing; five cover repetition and stability; and five cover resource use. Four of the forty are inert on this corpus, two constant and two duplicating another measure, leaving thirty-six live.

Attribution uses three fixed constants: a minimum detectable effect of $0.5$, a pooled $|t|$ threshold of $2.5$, and an effect gate of $0.25$. These follow from the generator's coefficient range rather than from any episode. Two scope conditions apply. The ladder is read from interventional contrasts only, so a candidate an agent resolved by regression on observational rows counts as never tested, which is why the toolbox arms of Section 5.3 are reported on outcome measures alongside it. And because several contrasts on one candidate are pooled by inverse variance, a later contrast that disagrees with an earlier one can move a candidate between the upper states; only the never-tested boundary is monotone.

A two-parameter logistic item-response fit (Embretson & Reise, 2000) over 13 models and 41 instances converges with ability correlating to raw score at $r = 0.96$; the 8 instances below a discrimination of $0.5$ are reported rather than dropped.

## H SAMPLER ROBUSTNESS

Repairing the sampler moves nothing, and we bound how much it could have moved. An earlier version of the environment drew population rows with replacement, which shrinks standard errors as $1/\sqrt{n}$ without adding information. We reran an entire $4 \times 4$ design under both samplers with everything else held fixed, in the regime where the artifact should bite hardest: no per-call cap and budgets to 125 times, where a single call can draw a large share of the population. Over 480 matched pairs the difference in exact recovery is $-0.006$, 95% bootstrap CI $[-0.033, +0.021]$ (Efron, 1979), with 433 of 480 episodes agreeing exactly, and no trend across the budget axis along which exposure rises. We therefore bound the artifact at four percentage points on the budget and tool conclusions, where exposure is highest; the remaining studies run under a per-call cap and are less exposed still.