WORLDHATCH: CAN SCIENTIFIC AGENTS SUPPORT THEIR CAUSAL CLAIMS WITH EXPERIMENTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Can scientific agents autonomously complete experiments that support their own causal claims, even when their models are correct? We introduce WORLDHATCH, a diagnostic benchmark whose main design holds a two-path structure fixed. Agents retain, revise, or expand a documented explanation. The design separates evidence that an influence exists, acts selectively, and disappears under isolation. We assess model correctness and successful operation separately from these measured comparisons. In the main evaluation of seven language models, the strongest agent correctly models and uses the additional influence in every episode where it exists, but completes all three evidence requirements in only 0.5%. We characterize these evidence gaps with executable alternatives allowed by the task that fit the recorded measurements while contradicting a causal claim in the submitted model. Across models in the main evaluation, verified continuations complete the missing evidence and another successful operation within budget for almost all otherwise successful but evidence-incomplete episodes with that influence, without changing their models or policies. In a matched prompt comparison, specifying all three verification targets without prescribing an experiment sequence yields near-complete evidence collection for one frontier model but leaves large gaps for two others. The correctness–evidence gap also appears in an asymmetric cross-path extension. Evaluation should distinguish correct modeling, guided verification, and autonomous evidence completion.

1 INTRODUCTION

Scientific agents based on large language models (LLMs) are increasingly being developed to formulate hypotheses, design experiments, interpret results, and guide further investigation across scientific domains (Zheng et al., 2025; Roohani et al., 2025; Gottweis et al., 2026). Scientific discovery can require looking beyond the causal scope of an existing explanation. When observations depart from expectations, revising relationships among known causes may be insufficient, and an overlooked influence may need to be incorporated. Such an expanded explanation must do more than fit the observations already collected. It also makes causal claims about how the system responds under other interventions, and these claims require experimental support (Imai et al., 2013). A model may predict these responses correctly while the agent's measurements remain compatible with competing causal explanations. The agent must therefore autonomously select and perform further experiments to distinguish among these explanations and rule out alternatives to its causal claims. Do agents autonomously select and complete the experiments needed to distinguish their evolving causal explanations from competing alternatives?

Model construction and experiment selection are distinct parts of scientific inquiry (Klahr & Dunbar, 1988; Hyttinen et al., 2013). Schauble et al. (1991) likewise distinguish an engineering goal of producing a desired outcome from a science goal of understanding causal relations. PhysGym examines supplied priors (Chen et al., 2025a), CausaLab separates predictive accuracy, mechanism recovery, and residual causal ambiguity (Yang et al., 2026), and ReplaySCM tests compatible alternatives and separating interventions (Batzoglou, 2026). Our evidence analysis examines the experiments of agents with exact models and successful control. We ask whether their measurements rule out alternatives to the causal claims encoded in their models, independently of an evaluator's verification of their answers. Our target is autonomous completion of experimental support for an agent's own claims, rather than experimental design with every verification target supplied in advance.

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.

We introduce WORLDHATCH, a diagnostic benchmark for evaluating whether agents autonomously complete experiments supporting their own causal claims. Its main design uses an auxiliary input, two separately controlled paths, and simultaneous output measurements. This structure makes source dependence, selectivity, and blocking separately testable through controlled comparisons (Chen & Klahr, 1999). Operation uses connected paths, but models also predict isolated responses, so correct modeling and successful operation need not include evidence for those predictions. Rather than introduce multiple complex structures, we vary response laws within this fixed design to compare evidence collection without changing the interventions or verification problem. We separately test asymmetric cross-path interactions with corresponding evidence criteria (Section 5.4).

The main benchmark comprises 216 interactive episodes across six physical and computational families. Within a fixed action budget, agents submit experimental and operational actions to a simulator, receive measurements and outcome feedback, and use these observations to revise their models and choose subsequent experiments. Three world variants share the prompt and controls. In RETAIN, the documented model is correct. In REVISE, a relation must change without adding a causal input. In EXPAND, the account must include the auxiliary source as a cause. We call this expansion *causal boundary discovery* (Figure 1(b)). With the source idle, REVISE and EXPAND produce identical discrepancies, so the initial anomaly cannot determine the appropriate revision. Their deterministic, stationary responses over finite input domains allow separate checks of model correctness, successful operation at previously untested inputs, and experimental evidence. Our primary metric, *Discovery Success*, combines correct modeling and successful operation with the required claim-specific experimental support. In EXPAND, this includes source dependence, selectivity, and blocking. Our evidence analysis tests whether the observations admit an executable alternative and whether further experiments can complete the evidence within budget while leaving the model and policy unchanged.

Our contributions are:

- A claim-specific evaluation of experimental support for causal models that separates establishing an additional influence, testing selective path effects, and testing blocking at the measured settings.
- WORLDHATCH, 216 executable episodes across six apparatus families, with a shared two-path design and separate checks of exact modeling, control, and experimental evidence.
- An evaluation of seven language models showing that correct modeling and successful operation can coexist with incomplete experimental support. Verified continuations complete the missing evidence within the remaining budget for almost all otherwise successful EXPAND episodes with incomplete evidence in the main evaluation, without changing their models or policies. An asymmetric extension reproduces the gap with different path claims.

2 CAUSAL CLAIMS AND EXPERIMENTAL EVIDENCE

In EXPAND, our evidence criteria target three structural claims of boundary expansion rather than every prediction at every setting. An input may affect a response, a connection control may selectively change an output, and isolation may remove the input's effect. The comparisons support these claims at measured settings, while exact model checking separately covers all legal inputs. Let x denote the inputs within the initial causal scope and s an available auxiliary input omitted from that scope. An intervention $q = (x, s)$ returns a response y under a deterministic, stationary law:

$$y = F(q), \qquad D_t = \{(q_i, y_i)\}_{i=1}^{t}. \tag{1}$$

Here D_t records the interventions measured so far. All controls and their operational meanings are public, but their effects on the responses must be investigated.

The first question is whether the model's causal scope needs to expand. The manual provides an initial model $m(x)$ independent of s. The agent can provisionally retain it while it remains compatible with the observations. If m fails but another model using only x still fits, a correction within the initial scope remains possible. In contrast, the controlled observations

$$x_i = x_j, \qquad s_i \neq s_j, \qquad y_i \neq y_j \tag{2}$$

cannot both be explained by any deterministic model independent of s (Figure 1(c)). This contrast establishes the effect of s regardless of how the model is represented. Including s as a cause is

causal boundary discovery. The contrast is sufficient, not necessary under every stronger structural prior. In RETAIN and REVISE, the evidence requirement is a matched source comparison with both paths connected and both outputs unchanged. This records invariance at tested settings, not global absence of an effect.

The second question is whether a proposed path has a selective effect. WORLDHATCH provides two separately controlled service paths and measures both outputs on every trial. Here $x = (u, c, a, b)$ comprises primary demand u, context c, and binary gates a, b, where 1 connects a path and 0 isolates it. The outputs are y_{L} and y_{R}. Separate control of the paths does not guarantee that each output is independent of the opposite gate. Holding u, c, b fixed, we write the outputs as $y_{\mathrm{L}}(a, s)$ and $y_{\mathrm{R}}(a, s)$. A certificate is a set of measurements satisfying the evidence conditions for the evaluated claims. For the left path, three measured settings must satisfy

$$\begin{aligned} &y_{\mathrm{L}}(1, s_0) \neq y_{\mathrm{L}}(1, s_1), \\ &y_{\mathrm{L}}(0, s_k) \neq y_{\mathrm{L}}(1, s_k), \qquad s_0 \neq s_1, \quad k \in \{0, 1\}. \\ &y_{\mathrm{R}}(0, s_k) = y_{\mathrm{R}}(1, s_k), \end{aligned} \tag{3}$$

The first comparison establishes source dependence of the left output. The gate comparison changes the left output but leaves the right unchanged and shares its connected measurement at s_k with the source comparison. The right-path certificate swaps the outputs and gates. The two certificates may use different source values and primary settings, with measurements collected in any order. Together they establish source dependence and selective effects on both paths.

The third question is whether isolation removes the source effect. Blocking is distinct from selectivity and requires varying the source under isolation. For the left path, *Discovery Success* extends Equation 3 to four measured settings satisfying

$$\begin{aligned} &y_{\mathrm{L}}(1, s_0) \neq y_{\mathrm{L}}(1, s_1), \\ &y_{\mathrm{L}}(0, s_0) = y_{\mathrm{L}}(0, s_1), \\ &y_{\mathrm{R}}(0, s_k) = y_{\mathrm{R}}(1, s_k) \quad \text{for } k = 0, 1. \end{aligned} \tag{4}$$

With u, c, b fixed and $s_0 \neq s_1$, the source contrast vanishes under isolation while the other output is unchanged by isolation at both source values. Swap outputs and gates for the right path. Returning to the connected idle baseline can reflect cancellation between source and gate effects rather than blocking. Varying the source under fixed isolation distinguishes these explanations (Appendix F.2).

Model correctness is a property of predictions over all legal inputs. Experimental support is a property of D_t under stated assumptions. The evaluator can confirm that a submitted model is exact even when the agent's observations also fit an alternative denying one of its causal claims. Structural assumptions may guide model construction, but an assumption not guaranteed by the task does not replace the measured comparisons required to support the evaluated claims.

3 THE WORLDHATCH BENCHMARK

3.1 EXPERIMENTAL STRUCTURE AND DESIGN RATIONALE

The main benchmark fixes the two-path structure while varying response laws and task settings, keeping interventions and claim-specific evidence criteria comparable across mechanisms (Figure 1). The auxiliary source permits dependence tests, separately operated connections permit selective-effect tests, and source variation under isolation permits blocking tests. Both outputs remain observable. Section 5.4 separately introduces cross-path interaction with corresponding changes to the claims and comparisons.

The inputs are primary demand u, context c, auxiliary setting s, and binary gates a, b, with $x = (u, c, a, b)$. Each gate connects (1) or isolates (0) its own service path without changing the other connection or editing either meter. Delivery records $a \times s$ and $b \times s$ describe command delivery. They do not guarantee that changing one gate leaves the other output unchanged or that isolation removes the source effect. Both raw outputs $y = (y_{\mathrm{L}}, y_{\mathrm{R}})$ remain observable. Primary demand and context can be held fixed for comparisons or varied for model-based operation.

In the main benchmark, each output follows

$$y_j(q) = m_j(u, c) \oplus \delta \oplus g_j h(s), \qquad g_{\mathrm{L}} = a, \quad g_{\mathrm{R}} = b. \tag{5}$$

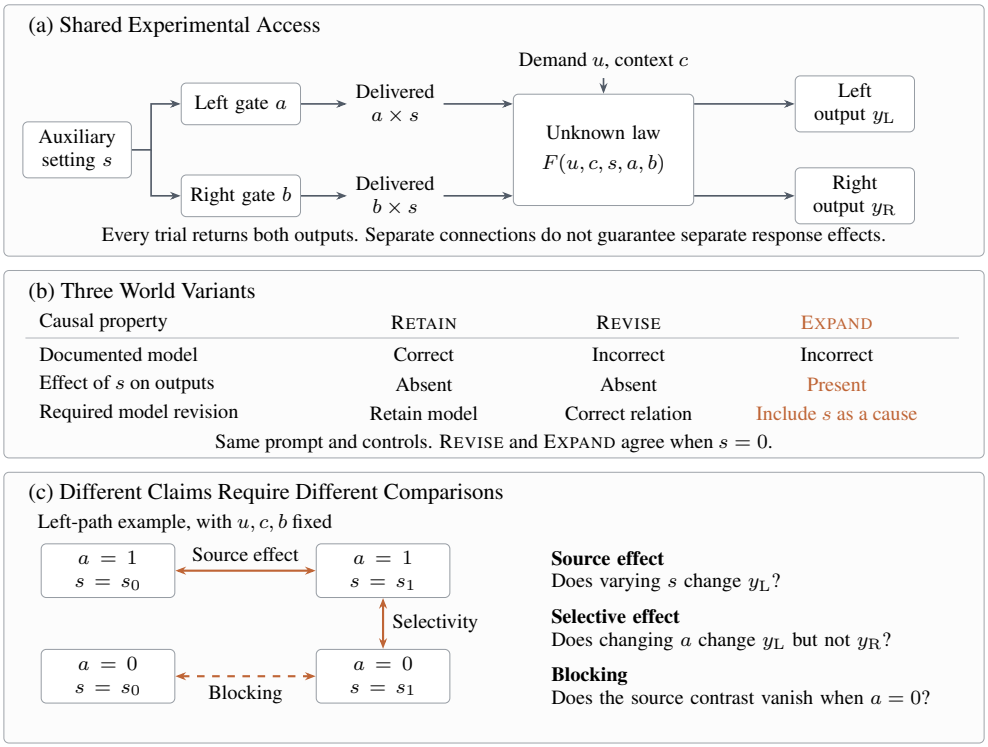


Figure 1: Experimental access and the claims it allows agents to test. (a) Connection controls and delivery records are public, while the response law must be inferred. (b) World variants change the needed revision without changing experimental access. (c) Three settings can establish a source effect and a selective gate effect. The right path has the corresponding comparisons.

Here m_j is the documented response, δ a local correction term, and $h(s)$ an auxiliary response transmitted through gate g_j. The operator $\oplus$ is addition, XOR, or channel rotation according to the apparatus. Agents receive the controls and manual, not this response structure. Their programs may express other deterministic laws within the task's program limits. Our evidence analysis searches these permitted functions for explanations compatible with the measurements, without assuming that agents know the structure of the response law used to generate the episodes.

Finite controls and stationary, deterministic responses make this separation verifiable. Every submitted program can be checked over its full input domain, and a competing program can be checked against every recorded measurement. We can also replay additional experiments without changing the law during an episode. The main release contains

$$\boxed{6 \text{ families } \times \text{ 4 mechanism variants } \times \text{ 3 narratives } \times \text{ 3 worlds} = 216 \text{ episodes}}.$$

The six families locate auxiliary influences in mechanical state, motion, measurement, computation, material flow, and command translation (Appendix B). The 24 mechanism configurations vary response laws within one shared topology, not across six independent causal structures. These comparisons concern finite, stationary systems without measurement noise or feedback. Source settings are $s \in \{0, 1, 2, 3\}$, with $s = 0$ idle. Parameters and unit names are shared across world and narrative variants.

3.2 WORLD AND NARRATIVE VARIANTS

The three world variants require retaining, revising, or expanding the documented model while preserving experimental access. They share the prompt, controls, objective, and budget. RETAIN uses $\delta = 0$ and $h \equiv 0$. REVISE has $\delta \neq 0$ and $h \equiv 0$. EXPAND shares that correction term but has nonconstant h with $h(0) = 0$, so $F_{\text{Expand}}(u, c, 0, a, b) = F_{\text{Revise}}(u, c, 0, a, b)$ for all admissible (u, c, a, b). Thus, idle-source experiments cannot distinguish these worlds. The same discrepancy

can therefore call for either a local correction or expansion to include s. Response functions remain fixed throughout interaction. Agents are not told which world variant they face.

These variants capture an investigator's uncertainty about whether an existing explanation should be retained, revised, or expanded. A proposed law must account for established behavior as well as responses under changed conditions, rather than merely fit the observations used to construct it. We therefore require controlled comparisons in all three worlds, testing the model's claims of invariance or causal influence at measured settings.

Narrative variants (skins) change task descriptions, not controls, parameters, or dynamics. Neutral describes the apparatus directly. Natural adds an application setting. Counter Prior adds a document claiming no auxiliary influence on the primary response, true in RETAIN and REVISE but false in EXPAND. These variants test sensitivity to narrative framing and prior claims under fixed dynamics.

3.3 INTERACTION AND PRODUCTION

At each turn, the agent submits an action and reason in JSON, optionally including theory and policy programs. `TRIAL` selects all five controls and returns both raw outputs together. Reading the other output requires no extra action, but obtaining a matched comparison may require another setting. `RUN` also specifies two corrections and attempts production. `INSPECT` reads the previous measurement, and `STOP` ends the episode. All actions, including invalid ones, consume the 20-turn budget. Interaction continues after production success until `STOP` or budget exhaustion.

The policy $\pi(q)$ selects two corrections before observing the outcome. The target pair for production remains $m(u, c)$ even when the manual incorrectly predicts the raw response. At every operational setting, where $a = b = 1$, the policy must satisfy

$$F(q) \oplus \pi(q) = m(u, c). \tag{6}$$

Production must also deliver the requested service code to both units. Model and policy programs are committed before the resulting measurement and read only public controls. Model checks cover every legal (u, c, s, a, b), and policy checks cover every (u, c, s) with $a = b = 1$, including $s = 0$. A `RUN` at previously untested inputs may use an idle source. For causal expansion, the prompt requires evidence distinguishing auxiliary effects from local errors and establishing reproducibility across the isolatable paths (Appendix D).

Under a declared structural prior, a reference policy achieves *Discovery Success* on all 216 episodes using six actions in RETAIN and REVISE and ten in EXPAND, including `STOP`, within the 20-action budget. Appendix D.5 gives its construction and verification.

4 EVALUATION PROTOCOL

4.1 AGENT SETUP

We evaluate Qwen3.8-27B-FP8, Qwen3.5-9B and 4B in BF16, GPT-OSS-20B in native MXFP4, and three frontier models via API, GPT-5.6-sol, GPT-6 Astra, and Gemini 3.8 Flash. Each has three runs on all 216 distinct episodes, with 72 per world. Pooled episode counts include repeats across runs. Episodes allow 20 actions and 4,096 output tokens per response, including reasoning. Unparseable responses receive one formatting retry at the same cap without new measurements. Appendix D.4 lists reasoning and decoding settings, and Appendix E.6 reports response-limit diagnostics.

4.2 METRICS

An autonomous scientific agent claiming a new causal discovery must, at a minimum, select and perform controlled interventions to test whether revising the existing explanation could account for the observations without adding a cause. In WORLDHATCH, an output contrast when only s changes rules out explanations confined to the initial causal scope, rather than merely showing that the documented model is wrong. Identifying this influence is only the first step. If the expanded model also predicts selective path effects and blocking, those claims require their own measured comparisons. The interface specifies command delivery, not these response properties.

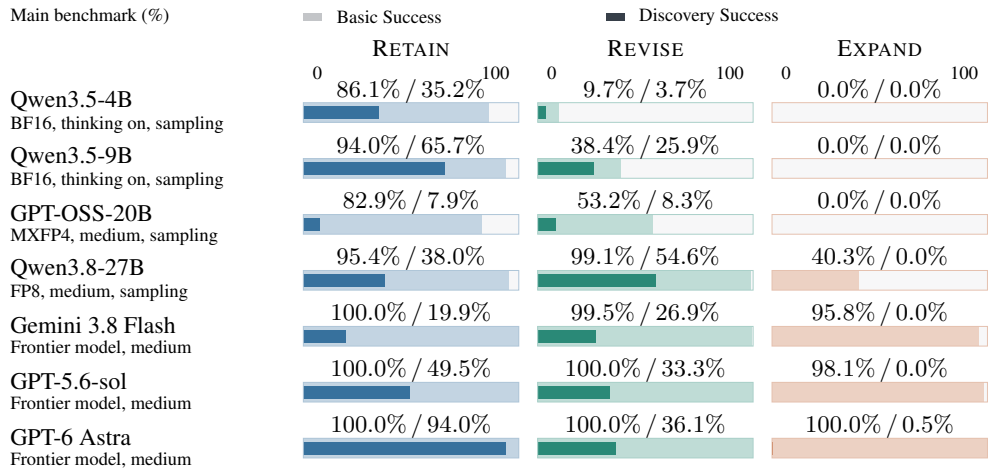


Figure 2: Main benchmark results (%). Values are means over three runs of 72 episodes per world. Light/dark bars and paired labels show *Basic Success* / *Discovery Success*, our primary metric.

Discovery Success evaluates this experimental responsibility by requiring correct modeling and successful operation together with controlled evidence for the model's causal claims evaluated by the benchmark. Scoring is deterministic and does not use an LLM judge. *Discovery Success* requires its evidence by a `RUN` that satisfies *Basic Success* (a qualifying `RUN`), including that operation's measurement.

- *Discovery Success*, our primary metric, adds evidence to *Basic Success* by the same qualifying `RUN`. In RETAIN and REVISE, two measurements vary only s with both gates connected ($a = b = 1$) and both outputs unchanged. In EXPAND, isolation must remove the source contrast and leave the other output unchanged at both source values, in addition to source and selectivity evidence on both paths (Section 2).
- *Basic Success* requires an exact model and a globally valid policy, submitted before a correctly predicted successful `RUN` at previously untested inputs. *Theory Accuracy* checks the model over all legal inputs. *Control Success* checks policy validity over all operational settings and a successful `RUN` using it. Unlike these component scores, *Basic Success* requires both at the same operation, without an experimental-evidence requirement. This definition is identical in RETAIN, REVISE, and EXPAND.

Detailed scoring rules are provided in Appendix D.2–D.3.

In RETAIN and REVISE, *Active Trials* counts episodes with a measured `TRIAL` at $s > 0$. *Matched Trials* counts episodes with fixed-u, c source comparisons at $a = b = 1$, irrespective of output. Both use all `TRIAL`s, whereas discovery scoring also accepts `RUN` measurements. For EXPAND episodes with *Basic Success*, we test whether a permitted alternative fits every observation while denying source dependence, a selective gate effect, or blocking (Appendix F). The asymmetric extension uses the same correctness and operation checks with structure-specific evidence criteria (Section 5.4).

5 RESULTS

5.1 MAIN RESULTS

Figure 2 compares *Basic Success* with our primary metric, *Discovery Success*. Submitted models make predictions even for interventions the agent never measured, including isolation settings. Discovery asks whether the agent collected evidence for these causal claims. The evaluator's exhaustive checks are not evidence collected by the agent.

In EXPAND, Astra, Sol, and Gemini achieve *Basic Success* of 100.0%, 98.1%, and 95.8%, respectively. Yet *Discovery Success* is only 0.5% for Astra and zero for every other model. All three frontier models usually construct and use the correct model without completing the experimental

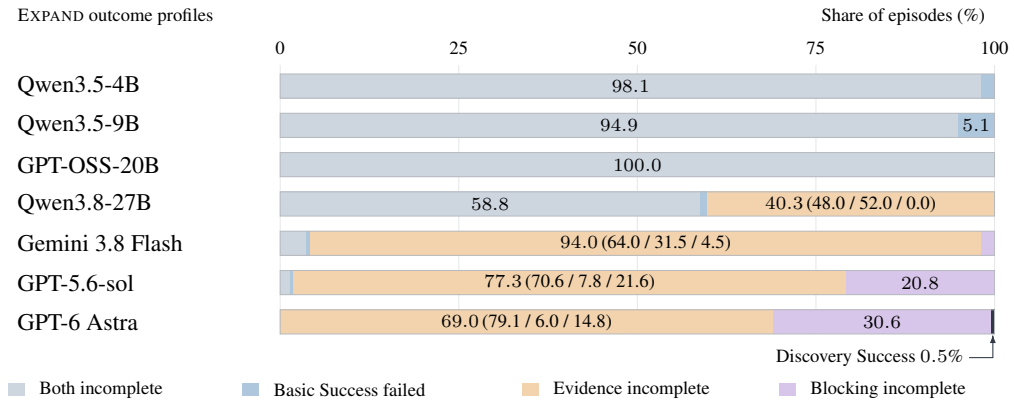


Figure 3: EXPAND modeling and causal-evidence outcomes (%). Values are means over three runs of 72 episodes per world. Bars partition episodes by *Basic Success* and completion of source, selectivity, and blocking evidence. Among evidence-incomplete episodes, the percentages in parentheses indicate matched comparisons for neither gate, one gate, or both gates, in that order.

support for its causal claims. In RETAIN and REVISE, Astra has perfect *Basic Success* but *Discovery Success* of 94.0% and 36.1%. Correctly retaining or correcting an explanation does not ensure that agents test its prediction that the auxiliary input leaves the outputs unchanged.

All local models have zero EXPAND *Discovery Success*. Qwen 27B nevertheless shows that evidence collection need not follow the same ranking as correct modeling and operation. In REVISE, Qwen 27B and Gemini have nearly identical *Basic Success* (99.1% and 99.5%), but *Discovery Success* is 54.6% and 26.9%, respectively. In EXPAND, Qwen 27B's 40.3% *Basic Success* coexists with zero *Discovery Success*. Smaller-model failures also involve response format, model construction, and control.

5.2 FAILURE ANALYSIS

Figure 3 distinguishes failure of *Basic Success* from missing source, selectivity, or blocking evidence. The 4B, 9B, and 20B models predominantly leave both incomplete, while Qwen 27B exhibits both failure patterns. For frontier models, *Evidence incomplete* accounts for 69.0–94.0% of all EXPAND episodes, including the 149 Astra episodes missing this evidence. To locate the missing evidence, we examine matched gate comparisons in the 606 evidence-incomplete episodes across four models (Qwen3.8-27B, Gemini 3.8 Flash, GPT-5.6-sol, and GPT-6 Astra). Each comparison changes one gate alone, holding u, c, s and the opposite gate fixed with $s > 0$. Coverage counts these pairs regardless of the measured outputs, rather than whether they establish a selective effect.

The key distinction is between trying an isolation setting and controlling a comparison. Most failures lack the comparisons needed to form path evidence, with 408 of 606 lacking a pair for either gate. Astra and Sol tried disconnecting each path in every analyzed failure, but often lacked comparisons that changed only one gate while holding all other inputs fixed. Most episodes in each frontier failure cohort lack any matched gate comparison, including 79.1% of Astra's on average across runs. Qwen 27B more often covers one gate but never both, while Sol has the largest *Both gates* fraction. Pairs alone need not show a selective effect or connect to a source comparison. In only 14 failures, both comparison types are present for each output but do not share the connected measurements needed to form path evidence (Appendix F.5). These missing comparisons leave specific causal claims unresolved. For each of the 606 episodes, we construct an executable alternative allowed by the task that fits all recorded measurements while denying source dependence or a selective-effect claim (Appendix F.1). These alternatives make the evidence gaps explicit rather than define an independent failure rate. Here, unresolved claims are relative to the response functions allowed by the task and its stated assumptions. Stronger structural premises can exclude alternatives without additional measurements, but this does not establish which premises an agent used.

Astra completes source and selectivity evidence by a qualifying `RUN` in 31.0% of EXPAND episodes, but almost none also establish blocking. The *Blocking incomplete* category comprises 115 episodes

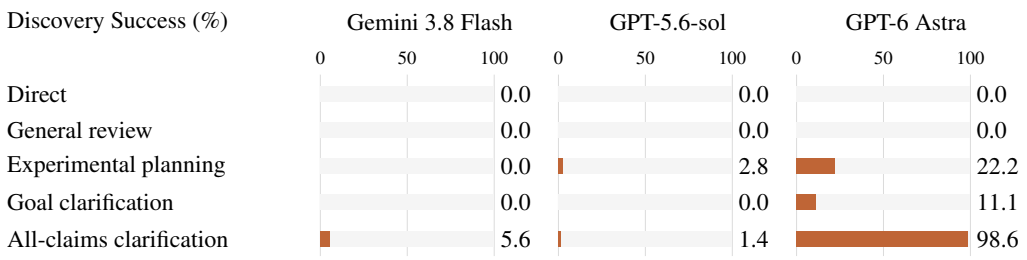


Figure 4: *Discovery Success* under five prompt conditions (%). Values are means over three runs of 24 Neutral EXPAND episodes.

with this gap across the four models. Each episode admits an executable alternative with a residual source effect under isolation that fits every measured output and retains every fully connected prediction (Appendix F.2). If such an alternative were the response law, the same recorded measurements would hold while the submitted model would predict an unmeasured isolation setting incorrectly. Successful production therefore does not resolve the model's claim that isolation removes the influence. Across both failure groups, comprising 721 episodes, we find alternatives with at most twice the submitted model's syntax-tree node count in 705 cases (97.8%). Each fits all recorded measurements while denying a causal claim in the submitted model (Appendix F.2). We also test whether these same episodes can complete the missing evidence within the remaining budget. Keeping the submitted model and policy fixed, we replace the final `STOP` with a continuation that completes source, selectivity, and blocking evidence and a successful, fully connected `RUN` at previously untested inputs. This is feasible within the remaining budget in 715/721 cases (99.2%). Replayed shortest continuations require 1–5 additional actions (median 3) (Appendix Table 33). These are constructed continuations, not agent actions. Section 5.3 tests whether guidance improves evidence collection.

5.3 PROMPT COMPARISON

Direct, the main benchmark prompt, tests autonomous identification and completion of evidence for the model's structural claims (Section 2). By design, agents must test response assumptions inferred from apparatus descriptions beyond the given command-delivery semantics (Section 3.1). This emphasizes scrutiny of apparatus assumptions and evidence-based model testing as practices of experimental science (Dounas-Frazer & Lewandowski, 2018; Holmes et al., 2015). In EXPAND, an exact submitted model predicts blocking under isolation, whether or not the prompt explicitly requests its verification. Explicit-goal conditions instead test completion of an external verification checklist, not autonomous identification of what needs testing.

Figure 4 compares five conditions on the same 24 Neutral EXPAND episodes, with three runs and identical budget caps. General review adds a completion check. Experimental planning supplies a matched-comparison strategy but leaves agents to identify which claims need testing. Goal clarification explicitly specifies source and selectivity goals and distinguishes command delivery from opposite-gate response independence. Relative to Direct, it changes both goals and interpretation of the control description. All-claims clarification retains that guidance and adds only the blocking goal. Neither prescribes comparisons or an action sequence (Appendix D.6).

All three models have zero *Discovery Success* under Direct and General review on this subset. Experimental planning yields 22.2%, 2.8%, and 0.0% for Astra, Sol, and Gemini, respectively. Even with a comparison strategy and an explicit pre-stop evidence check, agents rarely complete the required evidence when they must identify what needs testing themselves. By contrast, adding the blocking goal in All-claims clarification raises Astra's *Discovery Success* from 11.1% under Goal clarification to 98.6%. Astra can nearly complete verification when all targets are supplied, yet fails to complete the required experimental support reliably when it must identify those targets itself.

Explicit goals do not bring the same improvement for Sol and Gemini. All-claims clarification leaves Sol at 1.4% and Gemini at 5.6%. Some episodes compare measurements with an active source and an isolated path against those with an idle source and a connected path. Inferring blocking

from equal outputs in these settings requires the additional premise that changing the gate leaves the raw output unchanged when the source is idle. The command-delivery specification does not guarantee this response invariance. Without that premise, the comparison leaves gate and source changes confounded rather than directly testing source invariance under fixed isolation. The Sol and Gemini examples in Appendix D.7 explicitly describe restoring the idle baseline or testing whether isolation removes the effect. These are attempts to verify blocking, yet neither example varies the source with the relevant gate held isolated. The gap in these examples is therefore in applying the control-of-variables strategy to an intended test (Chen & Klahr, 1999), not an absence of relevant interventions. Other episodes vary the source under isolation but lack connected source contrasts at matching settings. Thus, explicit goals and relevant interventions need not yield complete controlled evidence.

5.4 Asymmetric Cross-Path Extension

We test whether correct modeling can coexist with incomplete evidence when the paths interact. Disconnecting the left path now reroutes the source, increasing its effect on the right output when that path is connected. *Discovery Success* now requires source dependence, an interaction between the gates that varies with the source, and blocking on each path. We evaluate three frontier models on 16 Neutral EXPAND instances across two families, with three runs each and the same action and response budgets. Agents receive a description of a shared supply system, but not the rerouting law or a list of verification targets. Further details are provided in Appendix G.

Table 1: Asymmetric cross-path extension (%). Three-run means and sample standard deviations over 16 instances per run.

Model	Basic Success	Discovery Success
Gemini 3.8 Flash	77.1 ± 3.6	2.1 ± 3.6
GPT-5.6-sol	91.7 ± 3.6	4.2 ± 3.6
GPT-6 Astra	100.0 ± 0.0	8.3 ± 3.6

Table 1 shows a large gap for all three models. Astra achieves *Basic Success* in all 48 episodes but *Discovery Success* in only four. Its models are exact over all 96 legal settings, including isolation, so the gap concerns experimental support for an already correct account of the cross-path effect. Sol and Gemini also rarely complete the evidence, even when their models and operation are correct.

All 122 otherwise successful but evidence-incomplete episodes admit an executable alternative that fits every recorded measurement while denying that the interaction between the gates changes with the source. With unchanged programs, replay-verified continuations complete the evidence and another successful `RUN` within the remaining budget in 105/122 cases (86.1%). The shortest require 1–5 additional actions (median 3). Thus, the gap also occurs in an asymmetric structure, where claim-specific comparisons, executable alternatives, and completion analysis apply to a different causal claim.

6 Conclusion

In this work, we test whether scientific agents autonomously complete experiments supporting their own causal claims, even when their models are correct. The deliberately fixed two-path design of WORLDHATCH separates evidence for the influence, its selective effects, and its blocking under isolation. Agents can model and use the influence correctly while leaving these claims unresolved by their measurements under the task's stated assumptions. Verified continuations show that almost all otherwise successful EXPAND episodes in the main evaluation could complete the evidence within budget using unchanged programs. Specifying all three verification goals nearly closes the evidence gap for Astra, but not Sol or Gemini. An asymmetric cross-path extension also separates exact modeling and successful operation from evidence for the model's path claims. Correct modeling, verification with explicit goals, and autonomous evidence completion are distinct achievements. Given only an environment and observations, an autonomous scientific agent must construct models, identify its own verification goals, and complete the experiments needed to support its claims.

AI USE STATEMENT

Coding agents assisted with implementation, testing, and project refactoring. Language models assisted with literature review, results analysis, manuscript refinement, prompt writing, and the natural-language presentation of dataset scenarios. Benchmark scoring used deterministic execution and recorded measurements, not an LLM judge. The authors are responsible for all text, code, data, and results.

ETHICS STATEMENT

WORLDHATCH is intended to support responsible scientific-agent development by evaluating whether agents collect experimental evidence for their causal claims. Its tasks concern measurement and control in synthetic environments, not the design or execution of harmful real-world experiments. The benchmark contains no personal data, involves no human-subject experiments, and does not operate physical laboratory equipment.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe the task interface and evaluation protocol. Appendices B–D and G document the response laws, prompts, model settings, and reference policies for both benchmark structures.

REFERENCES

John Alderete, Sebastian Benthall, Connie Xu, and John Xing. Separable Pathways for Causal Reasoning: How Architectural Scaffolding Enables Hypothesis-Space Restructuring in LLM Agents. *arXiv preprint arXiv:2604.20039*, 2026.

Serafim Batzoglou. ReplaySCM: A Benchmark for Executable Causal Mechanism Induction from Interventions. *arXiv preprint arXiv:2605.08197*, 2026.

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*, 2025a. Datasets and Benchmarks Track.

Zhe Chen and David Klahr. All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Development*, 1999.

Zhenhao Chen, Yongqiang Chen, Chenxi Liu, Junchi Yu, Xiangchen Song, Zijian Li, Jialin Li, Philip Torr, Bo Han, and Kun Zhang. CausalGame: Benchmarking Causal Thinking of LLM Agents in Games. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. In *International Conference on Learning Representations*, 2025b.

Davin Choo, Kirankumar Shiragur, and Arnab Bhattacharyya. Verification and Search Algorithms for Causal DAGs. In *Advances in Neural Information Processing Systems*, 2022.

Dimitri R. Dounas-Frazer and H. J. Lewandowski. The Modelling Framework for Experimental Physics: Description, Development, and Applications. *European Journal of Physics*, 2018.

Muhammad Qasim Elahi, Lai Wei, Murat Kocaoglu, and Mahsa Ghasemi. Adaptive Online Experimental Design for Causal Discovery. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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*, 2025.

Jiayi Geng, Howard Chen, Dilip Arumugam, and Thomas L. Griffiths. Are Large Language Models Reliable AI Scientists? Assessing Reverse-Engineering of Black-Box Systems. *arXiv preprint arXiv:2505.17968*, 2025.

Google DeepMind. Gemini 3.8 Flash Model Documentation, 2026. URL `https://ai.google.dev/gemini-api/docs/models/gemini-3.8-flash`. Accessed September 22, 2026.

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

Alex Havrilla, David Alvarez-Melis, and Nicolo Fusi. IGDA: Interactive Graph Discovery through Large Language Model Agents. *arXiv preprint arXiv:2502.17189*, 2025.

N. G. Holmes, Carl E. Wieman, and D. A. Bonn. Teaching Critical Thinking. *Proceedings of the National Academy of Sciences*, 2015.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candès, and Jure Leskovec. Automated Hypothesis Validation with Agentic Sequential Falsifications. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Antti Hyttinen, Frederick Eberhardt, and Patrik O. Hoyer. Experiment Selection for Causal Discovery. *Journal of Machine Learning Research*, 2013.

Kosuke Imai, Dustin Tingley, and Teppei Yamamoto. Experimental Designs for Identifying Causal Mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2013.

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*, 2024.

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*, 2026.

David Klahr and Kevin Dunbar. Dual Space Search during Scientific Reasoning. *Cognitive Science*, 1988.

Andrew Kyle Lampinen, Stephanie C. Y. Chan, Ishita Dasgupta, Andrew J. Nam, and Jane X. Wang. Passive Learning of Active Causal Strategies in Agents and Language Models. In *Advances in Neural Information Processing Systems*, 2023.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards Data-Driven Discovery with Large Language Models. In *International Conference on Learning Representations*, 2025.

Kevin Murphy. Model Discovery Agent: LLM-Assisted Bayesian Experiment Design for Data-Efficient Discovery of Mechanistic World Models. *arXiv preprint arXiv:2608.09696*, 2026.

OpenAI. GPT-5.6 Sol Model Documentation, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Accessed September 22, 2026.

OpenAI. GPT-6 Astra Model Documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Accessed September 22, 2026.

OpenAI. gpt-oss-20b Model Documentation, 2026c. URL https://developers.openai.com/api/docs/models/gpt-oss-20b. Accessed September 22, 2026.

Qwen Team. Qwen3.5-4B Model Card, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-4B. Accessed September 22, 2026.

Qwen Team. Qwen3.5-9B Model Card, 2026b. URL https://huggingface.co/Qwen/Qwen3.5-9B. Accessed September 22, 2026.

Qwen Team. Qwen3.8-27B-FP8 Model Card, 2026c. URL https://huggingface.co/Qwen/Qwen3.8-27B-FP8. Accessed September 22, 2026.

Martiño Ríos-García, Nawaf Alampara, Chandan Gupta, Indrajeet Mandal, Sajid Mannan, Ali Asghar Aghajani, N. M. Anoop Krishnan, and Kevin Maik Jablonka. AI Scientists Produce Results without Reasoning Scientifically. *arXiv preprint arXiv:2604.18805*, 2026.

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments. In *International Conference on Learning Representations*, 2025.

Mandana Samiei, Eunice Yiu, Anthony GX-Chen, Dongyan Lin, Jocelyn Shen, Blake A. Richards, Alison Gopnik, and Doina Precup. Human Adults and LLMs as Scientists: Who Benefits from Active Exploration? In *Proceedings of the 48th Annual Conference of the Cognitive Science Society*, 2026.

Leona Schauble, Leopold E. Klopfer, and Kalyani Raghavan. Students' Transition from an Engineering Model to a Science Model of Experimentation. *Journal of Research in Science Teaching*, 1991.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. LLM-SR: Scientific Equation Discovery via Programming with Large Language Models. In *International Conference on Learning Representations*, 2025.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D. Goodman. Hypothesis Search: Inductive Reasoning with Language Models. In *International Conference on Learning Representations*, 2024.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is Your Agent Smarter than a 5th Grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

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*, 2026.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The Causal-Neural Connection: Expressiveness, Learnability, and Inference. In *Advances in Neural Information Processing Systems*, 2021.

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. *arXiv preprint arXiv:2605.26029*, 2026.

Qing Zhao, Haowei Li, Weijian Deng, Pengxu Wei, and Liang Lin. EvoSCM: Scientific Belief Revision through Causal Model Evolution and Experimentation. *arXiv preprint arXiv:2609.01526*, 2026.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Tianshi Zheng, Kelvin Kiu-Wai Tam, Newt Hue-Nam K. Nguyen, Baixuan Xu, Zhaowei Wang, Jiayang Cheng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. NewtonBench: Benchmarking Generalizable Scientific Law Discovery in LLM Agents. In *International Conference on Learning Representations*, 2026.

APPENDIX CONTENTS

A Related Work **15**

B Executable Environments **16**

B.1 Response Functions 16

B.2 Examples across the Six Families 17

B.3 Skins and Public Information 19

C Public Problems and Worked Solutions **20**

C.1 Instructions Shared by the Six Problems 20

C.2 Family-Specific Problems and Solutions 21

C.3 World Variants: The Same Problem, Different Explanations 26

C.4 Matched Narrative Skins 27

D Evaluation Details **27**

D.1 Model and Action Space 27

D.2 Production at Untested Inputs 27

D.3 Intervention Evidence 28

D.4 Models and Inference Settings 28

D.5 Solvability Reference 29

D.6 Prompt Conditions 31

D.7 Evidence Completion and Operation Timing 32

E Additional Results **34**

E.1 Local Model Prompt Comparison 35

E.2 Outcome Breakdown 35

E.3 Family and Skin Results 36

E.4 Exploration without Boundary Expansion 37

E.5 Output-Budget Sensitivity 38

E.6 Frontier Response-Limit Diagnostics 38

F Evidence Analysis **40**

F.1 Evidence Sufficiency through Executable Alternatives 40

F.2 Analyzing Blocking Claims 41

F.3 A Correct Model with and without Matched Path Evidence 43

F.4 Completion within the Remaining Budget 44

F.5 Recorded Explanations and Comparisons 46

F.6 Family Differences in Comparison Completion 47

G Asymmetric Cross-Path Extension **48**

H Limitations and Discussion **51**

A Related Work

Experimental investigation distinguishes establishing an effect from identifying the process through which it arises. Imai et al. (2013) study how experimental designs and assumptions support identification of causal mechanisms. WorldHatch makes a related distinction operational in finite response systems with directly manipulated connection controls. Source dependence, selective gate effects, and complete blocking require different observed comparisons. These are claims at tested intervention settings, rather than estimates of natural direct or indirect effects. The two-path structure provides experimental access to each claim while holding the response law fixed during interaction.

Scientific-agent benchmarks cover data-driven discovery (Majumder et al., 2025; Chen et al., 2025b), interactive experiments (Wang et al., 2022; Jansen et al., 2024), law discovery (Zheng et al., 2026; Wiemann et al., 2026), and active experimentation (Gandhi et al., 2025; Geng et al., 2025). Methods support hypothesis testing and model search (Huang et al., 2025; Wang et al., 2024; Shojaee et al., 2025), or couple mechanism discovery with experiment selection (Kabra et al., 2026; Murphy, 2026). IGDA selects edge queries using model uncertainty and updates a predicted graph from binary feedback (Havrilla et al., 2025). In WorldHatch, agents instead receive measured responses and must compare them to establish causal claims.

CausaLab reports both high predictive accuracy with weaker mechanism recovery and residual causal ambiguity even when recovery is high (Yang et al., 2026). It also identifies premature commitment and reports that explicitly checking hypotheses against past observations improves task accuracy in a four-node setting. Our completion analysis goes beyond observing early commitment by computing the minimum additional actions needed for the specified evidence and a new successful operation while keeping the submitted programs fixed. Samiei et al. (2026) track information gain and the number of hypotheses compatible with successive observations in a blicket-style causal learning task. CausalGame uses LLM-judged causal-reasoning rubrics (Chen et al., 2026), while Corral analyzes evidence use through LLM-annotated epistemic graphs with expert validation (Ríos-García et al., 2026). ReplaySCM's Alternative-SCM task asks for an observation-compatible mechanism and an intervention separating it from a reference model. Its evidence checks identify alternatives and add observations to exclude them (Batzoglou, 2026). WorldHatch analyzes agent-collected evidence in episodes that already achieve exact modeling and successful control. It constructs executable alternatives that fit every measurement while denying a specific causal claim.

Model revision also concerns which hypotheses an agent can consider. EvoSCM maintains competing SCM hypotheses, selects discriminating experiments, and revises structures and mechanisms when predictions conflict with observations (Zhao et al., 2026). Alderete et al. (2026) study architectural support for restructuring the hypothesis space, including a condition in which the causal rule changes during an episode. WorldHatch holds response laws fixed and makes all controls available from the start. The agent must determine whether an initially omitted input affects the response, separating revision of an incomplete account from adaptation to a changing mechanism.

Causal identification distinguishes fitting data from determining a causal query under stated assumptions. Xia et al. (2021) study this distinction through neural causal models that agree on an observational distribution but may disagree on an interventional query. Our evidence analysis instead examines finite records of agent-selected interventions within a declared response language. Experiment design separates competing explanations under stated assumptions (Hyttinen et al., 2013; Elahi et al., 2024), and Choo et al. (2022) characterize intervention sets for verifying a proposed causal DAG while minimizing intervention costs. Our completion analysis starts from an agent's existing measurements and holds its submitted model and policy fixed. It finds the minimum additional actions needed to complete claim-specific evidence and a successful `RUN` at previously untested inputs, verifying each in-budget continuation by replay.

Prior work distinguishes hypothesis search from experiment selection (Klahr & Dunbar, 1988) and production from causal investigation (Schauble et al., 1991). PhysGym evaluates interactive physical discovery (Chen et al., 2025a). Prior work studies explicit control-of-variables instruction in humans (Chen & Klahr, 1999) and learning causal experimentation strategies from demonstrations and explanations in language models (Lampinen et al., 2023). Our matched-comparison diagnostic distinguishes performing an isolation trial from holding the other controls fixed to establish its effect. The prompt comparison then tests whether specifying claims and assumptions or supplying an experiment strategy improves the collection of that evidence.

B EXECUTABLE ENVIRONMENTS

Table 2 summarizes the six families and their measured responses. The fixed release contains 24 cells. Each combines one family and one source-effect variant, with the same instance constants shared across three worlds and three skins. The families share the two-path template in Equation 5, but implement its effects at different physical or computational nodes.

Table 2: Six insertion locations for an auxiliary influence.

Family	Natural setting	Effect location and measured response
C1	Greenhouse interlock	Pressure changes a mechanical contact read by a continuity meter.
C2	Packaging conveyor	Purge pressure contributes to measured shaft displacement.
C3	Fluorescence monitor	A beacon changes the reported detector address without moving the emitter.
C4	Metrology service	Maintenance traffic affects a worker operand and its numerical result.
C5	Microfluidic dosing	Wash fluid contributes to the volume reaching a collector.
C6	Laboratory cabinet	Maintenance messages affect command translation and the physical relay.

B.1 RESPONSE FUNCTIONS

The primary domain is $u \in \{0,1\}$ in C1 and C6 and $u \in \{0,1,2\}$ in C2–C5. All families use $c, a, b \in \{0,1\}$ and $s \in \{0,1,2,3\}$. The primary marker reports the selected value of u. It is not an independent measurement of the output. Each trial returns two raw measurements and service receipts (as, bs). The receipts are invariant across worlds and do not establish an effect on either raw response.

Equation 5 gives the common response form. The operator $\oplus$ is XOR for C1 and C6, addition modulo three for C3, and rational addition for C2, C4, and C5. Both components of the manual target pair are

$$m_j(u,c) = \begin{cases} u & \text{C1, C3, C5,} \\ 2u + c & \text{C2,} \\ k(u + o + c) & \text{C4,} \\ u \,\mathrm{XOR}\, c & \text{C6.} \end{cases}$$

The metrology scale k and offset o are public instance constants. The residual in REVISE and EXPAND is $\delta = 1$, except that the dosing residual is $1/2$. RETAIN uses $\delta = 0$. The auxiliary effect is zero outside EXPAND.

The four numeric source patterns are

$$s, \qquad s^2, \qquad \min(s,1), \qquad \mathbf{1}[s = 3].$$

C2 and C4 multiply the pattern by a generated gain in $\{1,2\}$. C5 uses that gain divided by two. C3 uses gain one or two and reduces the result modulo three. For the binary families, the active source sets are $\{1\}$, $\{2\}$, $\{1,2\}$, and $\{1,3\}$, respectively. These functions are private mechanism definitions, not answer templates supplied in the agent prompt. On the released instance parameters, the four EXPAND variants in each family have distinct source-effect tables over $s \in \{0,1,2,3\}$. This was checked using measured outputs relative to the idle response, with both paths connected. No such distinction is required in worlds without an auxiliary effect.

Table 3 lists every released source-effect table and the metrology constants. These values specify the evaluated instances. The catalog seed is 490731. It determines one instance seed for each family and variant, from which the scale, offset, gain, and unit names are drawn. Those assignments are reused across world variants and skins.

Table 3: Complete mechanism inventory. Values describe EXPAND. REVISE retains the same residual but uses $h = 0$, and RETAIN uses $\delta = h = 0$. Only C4 uses k, o.

Cell	k	o	δ	$(h(0), h(1), h(2), h(3))$
C1-R1	n/a	n/a	1	$(0, 1, 0, 0)$
C1-R2	n/a	n/a	1	$(0, 0, 1, 0)$
C1-R3	n/a	n/a	1	$(0, 1, 1, 0)$
C1-R4	n/a	n/a	1	$(0, 1, 0, 1)$
C2-R1	n/a	n/a	1	$(0, 2, 4, 6)$
C2-R2	n/a	n/a	1	$(0, 2, 8, 18)$
C2-R3	n/a	n/a	1	$(0, 1, 1, 1)$
C2-R4	n/a	n/a	1	$(0, 0, 0, 2)$
C3-R1	n/a	n/a	1	$(0, 2, 1, 0)$
C3-R2	n/a	n/a	1	$(0, 1, 1, 0)$
C3-R3	n/a	n/a	1	$(0, 1, 1, 1)$
C3-R4	n/a	n/a	1	$(0, 0, 0, 1)$
C4-R1	4	3	1	$(0, 1, 2, 3)$
C4-R2	4	5	1	$(0, 2, 8, 18)$
C4-R3	2	5	1	$(0, 2, 2, 2)$
C4-R4	4	7	1	$(0, 0, 0, 2)$
C5-R1	n/a	n/a	$\frac{1}{2}$	$(0, \frac{1}{2}, 1, \frac{3}{2})$
C5-R2	n/a	n/a	$\frac{1}{2}$	$(0, \frac{1}{2}, 2, \frac{9}{2})$
C5-R3	n/a	n/a	$\frac{1}{2}$	$(0, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$
C5-R4	n/a	n/a	$\frac{1}{2}$	$(0, 0, 0, 1)$
C6-R1	n/a	n/a	1	$(0, 1, 0, 0)$
C6-R2	n/a	n/a	1	$(0, 0, 1, 0)$
C6-R3	n/a	n/a	1	$(0, 1, 1, 0)$
C6-R4	n/a	n/a	1	$(0, 1, 0, 1)$

The six implementations place these effects at the nodes described in Table 2. To verify the simulator, we can independently alter the source emission or the state of the family-specific receiving node. These verification operations are unavailable to agents. The numerical response functions are also checked against an independently implemented arithmetic calculation.

B.2 EXAMPLES ACROSS THE SIX FAMILIES

Each example uses variant R1 from the fixed release. We describe the apparatus, compare observations that distinguish EXPAND from REVISE, and show how a fitted model supports an operation at previously untested inputs. In the equations below, $g_{\mathrm{L}} = a$ and $g_{\mathrm{R}} = b$ denote the two service connections. The four measurements in each example establish the source contrast and both three-setting path criteria. The complete reference schedules also probe the remaining source settings and the isolated idle settings before submitting a model and policy. Full problems and reference trajectories appear in Appendix C.

B.2.1 GREENHOUSE SHUTTER INTERLOCK

Two greenhouse vents have safety catches whose electrical continuity is measured separately. The primary command u sets the intended contact state. Cleaning pressure s is delivered through service sleeves controlled by a and b. In C1-R1, pressure setting 1 changes the contact position when its sleeve is connected. The EXPAND response is

$$y_j = u \,\mathrm{XOR}\, 1 \,\mathrm{XOR}\, \big(g_j \mathbf{1}[s = 1]\big).$$

At $u = 1, c = 0, s = 0, a = b = 1$, the meters read $(0, 0)$ instead of the manual's $(1, 1)$. The same discrepancy occurs in REVISE. Changing only s to 1 gives $(1, 1)$ in EXPAND, whereas REVISE remains at $(0, 0)$. At $s = 1$, closing the left sleeve gives $(0, 1)$ and closing the right sleeve gives $(1, 0)$. These comparisons concern measured contacts, not the recorded vent command or the service receipts.

A downstream polarity relay corrects the delivered signal using $\pi_j = 1 \text{ XOR } (g_j \mathbf{1}[s = 1])$. At the previously untested setting $u = 0, c = 1, s = 2, a = b = 1$, the model predicts raw contacts $(1, 1)$. Corrections $(1, 1)$ deliver the target $(0, 0)$, while both cleaning-service receipts equal 2.

B.2.2 PACKAGING CONVEYOR

Each lane has an encoder that measures shaft displacement after a conveyor stroke. A purge-air installation can be isolated independently from each lane. The manual predicts $2u + c$ millimetres. In C2-R1, the local offset is one millimetre and purge pressure contributes two millimetres per source unit through a connected branch:

$$y_j = 2u + c + 1 + 2g_j s.$$

At $u = 1, c = 0, s = 0, a = b = 1$, both encoders read 3 mm rather than the manual's 2 mm. This observation is shared by REVISE and EXPAND. Changing only s to 1 raises both readings to 5 mm in EXPAND, while REVISE stays at 3 mm. At $s = 1$, isolating the left branch gives $(3, 5)$ mm and isolating the right branch gives $(5, 3)$ mm. The purge installation therefore affects shaft motion, not just the service receipt.

A trim motor applies the precommitted correction $\pi_j = -(1 + 2g_j s)$. At the previously untested setting $u = 0, c = 1, s = 2, a = b = 1$, raw displacements are $(6, 6)$ mm. Corrections $(-5, -5)$ mm deliver the target $(1, 1)$ mm without disconnecting either purge branch.

B.2.3 FLUORESCENCE MONITOR

Two monitors report detector addresses for an illuminated LED at position u. A separate beacon sequencer communicates through service fibres controlled by shutters a and b. In C3-R1, the LED does not move, but the beacon changes the reported address:

$$y_j = (u + 1 + 2g_j s) \bmod 3.$$

At $u = 1, c = 0, s = 0, a = b = 1$, the monitors report $(2, 2)$ rather than the manual's $(1, 1)$. Both REVISE and EXPAND produce this mismatch. Changing only s to 1 gives $(1, 1)$ in EXPAND but leaves $(2, 2)$ in REVISE. At $s = 1$, closing the left service shutter gives $(2, 1)$ and closing the right gives $(1, 2)$. The separately recorded LED position remains 1. Thus the added influence acts on reporting, not emitter position.

The correction rotates each delivered address by $\pi_j = -(1 + 2g_j s) \bmod 3$. At the previously untested setting $u = 0, c = 1, s = 2, a = b = 1$, the raw reports are $(2, 2)$. Rotations $(1, 1)$ recover the target addresses $(0, 0)$ while retaining both beacon connections. The agreement with the manual at $s = 1$ alone would have concealed the two offsetting effects.

B.2.4 METROLOGY PROCESSING SERVICE

Two worker processes convert numerical measurements. A maintenance queue accepts request tags s, and a and b control the workers' service subscriptions. The C4-R1 manual uses scale 4 and offset 3, giving $m_j = 4(u + 3 + c)$. Before scaling, the worker operand receives the normalized perturbation $(1 + g_j s)/4$, combining the local offset and maintenance traffic. Multiplication by 4 gives the EXPAND output

$$y_j = 4(u + 3 + c) + 1 + g_j s.$$

At $u = 1, c = 0, s = 0, a = b = 1$, the results are $(17, 17)$ instead of the manual's $(16, 16)$ in both REVISE and EXPAND. Changing only the maintenance tag to $s = 1$ gives $(18, 18)$ in EXPAND, whereas REVISE remains at $(17, 17)$. At $s = 1$, disconnecting the left subscription gives $(17, 18)$ and disconnecting the right gives $(18, 17)$. The worker calculation still runs in each case, so the comparison tests the service's contribution to the numerical result.

The output correction is $\pi_j = -(1 + g_j s)$. At the previously untested setting $u = 0, c = 1, s = 2, a = b = 1$, the model predicts $(19, 19)$. Committed corrections $(-3, -3)$ deliver the manual target $(16, 16)$ while both workers receive the requested maintenance tag.

B.2.5 MICROFLUIDIC DOSING

The manual for each dosing channel is $m_j(u, c) = u$ microlitres. We illustrate the general setting with the first microfluidic variant in the fixed release. Its EXPAND response has a $0.5\,\mu$L local residual and a linear wash contribution:

$$V_{\mathrm{L}} = u + \tfrac{1}{2} + \tfrac{1}{2}as, \qquad V_{\mathrm{R}} = u + \tfrac{1}{2} + \tfrac{1}{2}bs.$$

Here $V_j = y_j$ is the raw collected volume in channel j. At $u = 1, c = 0, s = 0, a = b = 1$, each collector receives $1.5\,\mu$L rather than the commanded $1\,\mu$L. The REVISE counterpart produces exactly the same observation. Changing only s to 1 produces $(2, 2)$ in EXPAND but $(1.5, 1.5)$ in REVISE, in microlitres.

The EXPAND account can be tested further with one service valve closed. At $s = 1$, setting $a = 0, b = 1$ yields $(1.5, 2)$, while $a = 1, b = 0$ yields $(2, 1.5)$. Each valve changes only its own collector's volume relative to the connected measurement. Together with the source contrast, these four observations meet both three-setting path criteria. For production, a correct policy is

$$\pi_{\mathrm{L}} = -\tfrac{1}{2}(1 + as), \qquad \pi_{\mathrm{R}} = -\tfrac{1}{2}(1 + bs).$$

At the previously untested setting $u = 2, c = 1, s = 1, a = b = 1$, the model predicts raw volumes $(3, 3)$ and the `RUN` commits withdrawals $(-1, -1)$, delivering the primary target $(2, 2)$ in microlitres. The service receipts also equal $(1, 1)$, as requested. The complete reference schedule also measures wash settings 2 and 3 before fitting its model.

B.2.6 AUTOMATED LABORATORY CABINET

Each cabinet stores an instrument request and separately displays the physical relay state. The local condition c controls request polarity, so the manual predicts u XOR c. Maintenance notes arrive through subscriptions a and b. In C6-R1, note code 1 changes command translation and consequently the relay:

$$y_j = u \,\mathrm{XOR}\, c \,\mathrm{XOR}\, 1 \,\mathrm{XOR}\, \big(g_j \mathbf{1}[s = 1]\big).$$

At $u = 1, c = 0, s = 0, a = b = 1$, the relay lamps show $(0, 0)$ instead of the manual's $(1, 1)$ in both REVISE and EXPAND. Changing only the note code to $s = 1$ gives $(1, 1)$ in EXPAND but leaves $(0, 0)$ in REVISE. At $s = 1$, disconnecting the left subscription gives $(0, 1)$ and disconnecting the right gives $(1, 0)$. These are changes in the physical relays, not edits to the stored request text.

A hardware polarity actuator uses $\pi_j = 1$ XOR $(g_j \mathbf{1}[s = 1])$. At the previously untested setting $u = 0, c = 1, s = 2, a = b = 1$, raw relay states are $(0, 0)$. Corrections $(1, 1)$ deliver the target $(1, 1)$ while both supervisors receive note code 2. Unlike C1's mechanical contact effect, this influence enters through command translation before the relay acts.

B.3 SKINS AND PUBLIC INFORMATION

Natural narratives use the settings listed in Table 2. Neutral narratives describe the same instrument functions directly. Counter Prior narratives introduce a fallible certificate asserting auxiliary isolation. All renderings retain the same JSON controls, names, constants, and physical measurements within a matched cell. The private mechanism and world label never appear in the public episode identifier.

The initial packet provides the apparatus description, manual program, control meanings, action grammar, expression language, budget, and evaluation instructions. It does not contain a preselected causal model for EXPAND or a required sequence of experiments. The public service receipt makes the production requirement observable. No hidden evidence flag is used to decide whether a physical operation succeeds.

The agent receives only its selected public prompt and initial observation. Private world labels, mechanism records, instance seeds, and reference solutions are reserved for verification and excluded from the conversation.

C PUBLIC PROBLEMS AND WORKED SOLUTIONS

We select variant R1 in each of the six families, using the Natural skin. Each public problem is identical across its RETAIN, REVISE, and EXPAND counterparts. We show the family-specific prefix and print the common instructions once in Appendix C.1. Concatenating the prefix and those instructions reconstructs the complete initial prompt. Only display line breaks and JSON whitespace are reformatted. The initial observation appended by the runner is

```
Initial observation: {"meter": null, "turns_used": 0, "turns_remaining": 20}
```

The starting packet contains no trial outcome. The private explanations and reference trajectories below are not included in the agent conversation.

C.1 INSTRUCTIONS SHARED BY THE SIX PROBLEMS

The following is the shared suffix of the actual model-facing prompt. It states the target, program format, requirement to operate at previously untested inputs, and experimental-evidence objective without prescribing an experiment sequence. It explicitly requests evidence for expansion but does not prescribe invariant comparisons for retention or correction. Those comparisons evaluate measured support for the model's predicted invariance at tested settings (Section 2), rather than compliance with an enumerated test.

```
There are at most twenty actions. The response law remains fixed throughout
this episode. Keep the preceding history. Return one JSON object with action,
a nonempty concise reason, and optional theory and policy programs. No other
tools or private files are available.

Actions: {"op":"TRIAL","u":N,"context":C,"source":S,"a":A,"b":B} performs one
simultaneous physical trial with two independently metered outputs.
{"op":"RUN","u":N,"context":C,"source":S,"a":1,"b":1,
"correction":{"left":L,"right":R}} performs a production stroke with YOUR
precommitted corrections. The target for each specimen is the manual response
for u and context; the service target at each specimen is source. Both actual
endpoints are measured. The device never chooses corrections for you. For
numeric outputs delivered=raw+correction with |correction| <=32. Binary relays
use XOR and detector addresses use addition modulo three. For fluid, an
excessive withdrawal fails rather than clamping delivery to zero.
{"op":"INSPECT"} reads the last meter without actuation. {"op":"STOP"} ends
interaction. Every action, valid or invalid, costs one turn.

A theory predicts RAW left/right measurements. A policy outputs left/right
corrections. Programs are objects {"left":EXPR,"right":EXPR}. Expressions are
integers, rational strings, {"var":"NAME"}, or {"op":"OP","args":[...]}. Only
u, context, source, a and b are inputs. Operators: add, sub, mul, div, mod,
eq, lt, xor, and, or, not, if, min, max. Binary operators take two arguments,
not takes one and if takes condition/yes/no. Limits are 512 nodes, depth 20,
64-character literals, and numerator magnitude/denominator at most 10^12.
Values must be defined on every legal setting; relay values are binary and
channel values are 0..2. Models cannot read current outcomes, time or private
state. The manual and a zero-correction policy are initially installed. New
programs are committed BEFORE the selected action, without resetting any
apparatus. Wrong program syntax does not invalidate an otherwise
legal physical action.

Evaluation checks a submitted theory on EVERY legal intervention tuple, and a
policy on every operational tuple. No particular formula or hidden source-code
identity is required. After submitting a theory, validate it in a production
RUN at a setting that has not been measured in any earlier TRIAL or RUN in
this episode. A setting is the complete tuple (u, context, source, a, b). The
tuple must be new, not each individual value. Changing only the operation from
TRIAL to RUN, or changing only the correction, does not make a setting new.
You must choose this untested setting yourself; the environment will not issue
a separate validation task. Commit the theory and correction policy before
observing this RUN's outcome. The committed theory must correctly predict both
raw measurements, and the SAME RUN must achieve both production targets using
```

```
the committed policy. A correct prediction in a separate TRIAL followed by a
RUN at a previously tested setting does not satisfy this requirement. Physical
control alone is not a discovery certificate. Evidence for extending the
manual's causal scope must distinguish an auxiliary effect from an error in
the primary relation and establish reproducibility across the two
independently isolatable paths. Correct retention is allowed when the
measurements do not require expansion or correction. No scores or evaluator
event labels are returned during interaction.
```

C.2 Family-Specific Problems and Solutions

C1: Greenhouse shutter interlock

Public problem. This prefix is followed by Appendix C.1.

```
Two greenhouse vents have independently metered safety catches. The gardener
can command the vent position and operate a cleaning manifold. Each catch has
a service sleeve that can be sealed while the vent is tested. A downstream
polarity relay can change the delivered enable signal without
moving the catch.

Specimens left and right are named Birch and Willow. The primary manual
predicts each contact continuity in binary by: {"left": {"var": "u"}, "right":
{"var": "u"}} Its account uses the primary demand u and local operating
condition context. It does not give a primary-response law for the auxiliary
installation. Each trial starts the primary specimen at the same unloaded
mechanical or job condition. The auxiliary installation is not reset. The
recorded primary marker is the selected u, not a second measurement of the
response. This marker is especially distinct from the detector channel and
from the supervisory relay.

Controls have these meanings: u in [0, 1] sets the primary demand. For
context, 0 leaves the contact fixture at room temperature; 1 engages its
warming jacket. source in [0,1,2,3] means select the cleaning pressure
setting; zero is the idle auxiliary setting. a=0 means close the pneumatic
service sleeve for left; a=1 reconnects it. b does the same independently for
right. These controls do not remove a primary specimen or edit its meter. A
separate service receipt records a*source at left and b*source at right:
isolation prevents the selected service command from reaching that specimen. A
production RUN must deliver BOTH the primary target and the requested
auxiliary service code at both specimens. Calibration with isolated sleeves is
permitted and still returns ordinary measured responses.
```

Private worked solution. The following observations come from the released EXPAND reference policy. The fitted source effect at $s = (0, 1, 2, 3)$ is $(0, 1, 0, 0)$. The model and policy use Equations 8 and 9. Table 4 lists every reference action and raw measurement.

Table 4: C1-R1 Natural EXPAND reference. Outputs use the public units above. The final stopping action has no measurement.

Turn	Action	(u, c, s, a, b)	Raw outputs	Correction
1	TRIAL	$(1, 0, 0, 1, 1)$	$(0, 0)$	n/a
2	TRIAL	$(1, 0, 1, 1, 1)$	$(1, 1)$	n/a
3	TRIAL	$(1, 0, 2, 1, 1)$	$(0, 0)$	n/a
4	TRIAL	$(1, 0, 3, 1, 1)$	$(0, 0)$	n/a
5	TRIAL	$(1, 0, 0, 0, 1)$	$(0, 0)$	n/a
6	TRIAL	$(1, 0, 1, 0, 1)$	$(0, 1)$	n/a
7	TRIAL	$(1, 0, 0, 1, 0)$	$(0, 0)$	n/a
8	TRIAL	$(1, 0, 1, 1, 0)$	$(1, 0)$	n/a
9	RUN	$(0, 1, 1, 1, 1)$	$(0, 0)$	$(0, 0)$
10	STOP	n/a	n/a	n/a

At the previously untested production input, the committed model predicts $(0, 0)$ and the selected corrections deliver $(0, 0)$. The service receipt pair is $(1, 1)$. The reference completes source, selectivity, and blocking comparisons before this operation.

C2: PACKAGING CONVEYOR COMMISSIONING BENCH

Public problem. This prefix is followed by Appendix C.1.

```
You commission two packaging lanes. An encoder measures each completed
conveyor stroke. The facility also has an air-purge installation and an
isolator on each lane service branch. A trim motor can add or subtract a
preselected displacement during a stroke. An encoder reading arrives only
after that command is fixed.

Specimens left and right are named Birch and Alder. The primary manual
predicts each integrated shaft displacement in mm by: {"left": {"op": "add",
"args": [{"op": "mul", "args": [2, {"var": "u"}]}, {"var": "context"}]},
"right": {"op": "add", "args": [{"op": "mul", "args": [2, {"var": "u"}]},
{"var": "context"}]}} Its account uses the primary demand u and local
operating condition context. It does not give a primary-response law for the
auxiliary installation. Each trial starts the primary specimen at the same
unloaded mechanical or job condition. The auxiliary installation is not reset.
The recorded primary marker is the selected u, not a second measurement of the
response. This marker is especially distinct from the detector channel and
from the supervisory relay.

Controls have these meanings: u in [0, 1, 2] sets the primary demand. For
context, 0 uses the zero-offset stop; 1 adds a one-millimetre mechanical
indexing offset. source in [0,1,2,3] means select the purge-pressure demand;
zero is the idle auxiliary setting. a=0 means isolate the service-air branch
for left; a=1 reconnects it. b does the same independently for right. These
controls do not remove a primary specimen or edit its meter. A separate
service receipt records a*source at left and b*source at right: isolation
prevents the selected service command from reaching that specimen. A
production RUN must deliver BOTH the primary target and the requested
auxiliary service code at both specimens. Calibration with isolated sleeves is
permitted and still returns ordinary measured responses.
```

Private worked solution. The following observations come from the released EXPAND reference policy. The fitted source effect at $s = (0, 1, 2, 3)$ is $(0, 2, 4, 6)$. The model and policy use Equations 8 and 9. Table 5 lists every reference action and raw measurement.

Table 5: C2-R1 Natural EXPAND reference. Outputs use the public units above. The final stopping action has no measurement.

Turn	Action	(u, c, s, a, b)	Raw outputs	Correction
1	TRIAL	$(1, 0, 0, 1, 1)$	$(3, 3)$	n/a
2	TRIAL	$(1, 0, 1, 1, 1)$	$(5, 5)$	n/a
3	TRIAL	$(1, 0, 2, 1, 1)$	$(7, 7)$	n/a
4	TRIAL	$(1, 0, 3, 1, 1)$	$(9, 9)$	n/a
5	TRIAL	$(1, 0, 0, 0, 1)$	$(3, 3)$	n/a
6	TRIAL	$(1, 0, 1, 0, 1)$	$(3, 5)$	n/a
7	TRIAL	$(1, 0, 0, 1, 0)$	$(3, 3)$	n/a
8	TRIAL	$(1, 0, 1, 1, 0)$	$(5, 3)$	n/a
9	RUN	$(0, 1, 1, 1, 1)$	$(4, 4)$	$(-3, -3)$
10	STOP	n/a	n/a	n/a

At the previously untested production input, the committed model predicts $(4, 4)$ and the selected corrections deliver $(1, 1)$. The service receipt pair is $(1, 1)$. The reference completes source, selectivity, and blocking comparisons before this operation.

C3: FLUORESCENCE MONITOR RACK

Public problem. This prefix is followed by Appendix C.1.

```
Two fluorescence monitors view physically fixed LEDs at positions zero, one
and two. A facility beacon sequencer can be stimulated separately. Each
monitor has a shutter on its service fibre. The illuminated LED position is
read independently of the reported detector address. A preselected address
rotation acts on the delivered report, not on the LED.

Specimens left and right are named Cedar and Maple. The primary manual
predicts each reported detector channel in channel 0..2 by: {"left": {"var":
"u"}, "right": {"var": "u"}} Its account uses the primary demand u and local
operating condition context. It does not give a primary-response law for the
auxiliary installation. Each trial starts the primary specimen at the same
unloaded mechanical or job condition. The auxiliary installation is not reset.
The recorded primary marker is the selected u, not a second measurement of the
response. This marker is especially distinct from the detector channel and
from the supervisory relay.

Controls have these meanings: u in [0, 1, 2] sets the primary demand. For
context, 0 uses low reference illumination; 1 uses high reference
illumination. source in [0,1,2,3] means select the beacon pulse code; zero is
the idle auxiliary setting. a=0 means close the service-fibre shutter for
left; a=1 reconnects it. b does the same independently for right. These
controls do not remove a primary specimen or edit its meter. A separate
service receipt records a*source at left and b*source at right: isolation
prevents the selected service command from reaching that specimen. A
production RUN must deliver BOTH the primary target and the requested
auxiliary service code at both specimens. Calibration with isolated sleeves is
permitted and still returns ordinary measured responses.
```

Private worked solution. The following observations come from the released EXPAND reference policy. The fitted source effect at $s = (0, 1, 2, 3)$ is $(0, 2, 1, 0)$. The model and policy use Equations 8 and 9. Table 6 lists every reference action and raw measurement.

Table 6: C3-R1 Natural EXPAND reference. Outputs use the public units above. The final stopping action has no measurement.

Turn	Action	(u, c, s, a, b)	Raw outputs	Correction
1	TRIAL	$(1, 0, 0, 1, 1)$	$(2, 2)$	n/a
2	TRIAL	$(1, 0, 1, 1, 1)$	$(1, 1)$	n/a
3	TRIAL	$(1, 0, 2, 1, 1)$	$(0, 0)$	n/a
4	TRIAL	$(1, 0, 3, 1, 1)$	$(2, 2)$	n/a
5	TRIAL	$(1, 0, 0, 0, 1)$	$(2, 2)$	n/a
6	TRIAL	$(1, 0, 1, 0, 1)$	$(2, 1)$	n/a
7	TRIAL	$(1, 0, 0, 1, 0)$	$(2, 2)$	n/a
8	TRIAL	$(1, 0, 1, 1, 0)$	$(1, 2)$	n/a
9	RUN	$(0, 1, 1, 1, 1)$	$(0, 0)$	$(0, 0)$
10	STOP	n/a	n/a	n/a

At the previously untested production input, the committed model predicts $(0, 0)$ and the selected corrections deliver $(0, 0)$. The service receipt pair is $(1, 1)$. The reference completes source, selectivity, and blocking comparisons before this operation.

C4: METROLOGY PROCESSING SERVICE

Public problem. This prefix is followed by Appendix C.1.

```
A metrology service converts measurements in two independently instrumented
worker processes. A maintenance queue accepts tagged messages. The operator
can disconnect either worker service subscription while continuing a test
```

```
calculation. An output correction is committed before dispatch, so it cannot
copy the completed result.

Specimens left and right are named Hazel and Birch. The primary manual
predicts each completed numerical result in converted units by: {"left":
{"op": "mul", "args": [4, {"op": "add", "args": [{"op": "add", "args":
[{"var": "u"}, 3]}, {"var": "context"}]}]}, "right": {"op": "mul", "args": [4,
{"op": "add", "args": [{"op": "add", "args": [{"var": "u"}, 3]}, {"var":
"context"}]}]}} Its account uses the primary demand u and local operating
condition context. It does not give a primary-response law for the auxiliary
installation. Each trial starts the primary specimen at the same unloaded
mechanical or job condition. The auxiliary installation is not reset. The
recorded primary marker is the selected u, not a second measurement of the
response. This marker is especially distinct from the detector channel and
from the supervisory relay.

Controls have these meanings: u in [0, 1, 2] sets the primary demand. For
context, 0 applies the ordinary conversion; 1 adds one input unit of local
calibration offset before scaling. source in [0,1,2,3] means select the
maintenance request tag; zero is the idle auxiliary setting. a=0 means
disconnect the worker service-queue subscription for left; a=1 reconnects it.
b does the same independently for right. These controls do not remove a
primary specimen or edit its meter. A separate service receipt records
a*source at left and b*source at right: isolation prevents the selected
service command from reaching that specimen. A production RUN must deliver
BOTH the primary target and the requested auxiliary service code at both
specimens. Calibration with isolated sleeves is permitted and still returns
ordinary measured responses.
```

Private worked solution. The following observations come from the released EXPAND reference policy. The fitted source effect at $s = (0, 1, 2, 3)$ is $(0, 1, 2, 3)$. The model and policy use Equations 8 and 9. Table 7 lists every reference action and raw measurement.

Table 7: C4-R1 Natural EXPAND reference. Outputs use the public units above. The final stopping action has no measurement.

Turn	Action	(u, c, s, a, b)	Raw outputs	Correction
1	TRIAL	$(1, 0, 0, 1, 1)$	$(17, 17)$	n/a
2	TRIAL	$(1, 0, 1, 1, 1)$	$(18, 18)$	n/a
3	TRIAL	$(1, 0, 2, 1, 1)$	$(19, 19)$	n/a
4	TRIAL	$(1, 0, 3, 1, 1)$	$(20, 20)$	n/a
5	TRIAL	$(1, 0, 0, 0, 1)$	$(17, 17)$	n/a
6	TRIAL	$(1, 0, 1, 0, 1)$	$(17, 18)$	n/a
7	TRIAL	$(1, 0, 0, 1, 0)$	$(17, 17)$	n/a
8	TRIAL	$(1, 0, 1, 1, 0)$	$(18, 17)$	n/a
9	RUN	$(0, 1, 1, 1, 1)$	$(18, 18)$	$(-2, -2)$
10	STOP	n/a	n/a	n/a

At the previously untested production input, the committed model predicts $(18, 18)$ and the selected corrections deliver $(16, 16)$. The service receipt pair is $(1, 1)$. The reference completes source, selectivity, and blocking comparisons before this operation.

C5: MICROFLUIDIC DOSING BENCH

Public problem. This prefix is followed by Appendix C.1.

```
Two synthesis channels deliver a requested dose into separate graduated
collectors. The laboratory wash manifold can receive a separate bolus command.
Service-fluid valves are independently closable during calibration. Primed
trim reservoirs add or withdraw a quantity chosen before collection.
Withdrawal cannot remove fluid that is not present.
```

```
Specimens left and right are named Hazel and Willow. The primary manual
predicts each collected volume in microlitre by: {"left": {"var": "u"},
"right": {"var": "u"}} Its account uses the primary demand u and local
operating condition context. It does not give a primary-response law for the
auxiliary installation. Each trial starts the primary specimen at the same
unloaded mechanical or job condition. The auxiliary installation is not reset.
The recorded primary marker is the selected u, not a second measurement of the
response. This marker is especially distinct from the detector channel and
from the supervisory relay.

Controls have these meanings: u in [0, 1, 2] sets the primary demand. For
context, 0 uses neutral carrier fluid; 1 uses acidic carrier fluid. source in
[0,1,2,3] means select the wash bolus setting; zero is the idle auxiliary
setting. a=0 means close the service-fluid branch for left; a=1 reconnects it.
b does the same independently for right. These controls do not remove a
primary specimen or edit its meter. A separate service receipt records
a*source at left and b*source at right: isolation prevents the selected
service command from reaching that specimen. A production RUN must deliver
BOTH the primary target and the requested auxiliary service code at both
specimens. Calibration with isolated sleeves is permitted and still returns
ordinary measured responses.
```

Private worked solution. The following observations come from the released EXPAND reference policy. The fitted source effect at $s = (0, 1, 2, 3)$ is $(0, \frac{1}{2}, 1, \frac{3}{2})$. The model and policy use Equations 8 and 9. Table 8 lists every reference action and raw measurement.

Table 8: C5-R1 Natural EXPAND reference. Outputs use the public units above. The final stopping action has no measurement.

Turn	Action	(u, c, s, a, b)	Raw outputs	Correction
1	TRIAL	$(1, 0, 0, 1, 1)$	$(\frac{3}{2}, \frac{3}{2})$	n/a
2	TRIAL	$(1, 0, 1, 1, 1)$	$(2, 2)$	n/a
3	TRIAL	$(1, 0, 2, 1, 1)$	$(\frac{5}{2}, \frac{5}{2})$	n/a
4	TRIAL	$(1, 0, 3, 1, 1)$	$(3, 3)$	n/a
5	TRIAL	$(1, 0, 0, 0, 1)$	$(\frac{3}{2}, \frac{3}{2})$	n/a
6	TRIAL	$(1, 0, 1, 0, 1)$	$(\frac{3}{2}, 2)$	n/a
7	TRIAL	$(1, 0, 0, 1, 0)$	$(\frac{3}{2}, \frac{3}{2})$	n/a
8	TRIAL	$(1, 0, 1, 1, 0)$	$(2, \frac{3}{2})$	n/a
9	RUN	$(0, 1, 1, 1, 1)$	$(1, 1)$	$(-1, -1)$
10	STOP	n/a	n/a	n/a

At the previously untested production input, the committed model predicts $(1, 1)$ and the selected corrections deliver $(0, 0)$. The service receipt pair is $(1, 1)$. The reference completes source, selectivity, and blocking comparisons before this operation.

C6: AUTOMATED LABORATORY CABINET

Public problem. This prefix is followed by Appendix C.1.

```
Two laboratory cabinets store the requested instrument state and display the
actual relay lamp separately. A maintenance console accepts coded notes. Each
supervisor has a service-message subscription that can be disabled during
calibration. A hardware polarity actuator can correct the relay command, but
editing stored text does not move the physical relay.

Specimens left and right are named Alder and Cedar. The primary manual
predicts each physical relay state in binary by: {"left": {"op": "xor",
"args": [{"var": "u"}, {"var": "context"}]}, "right": {"op": "xor", "args":
[{"var": "u"}, {"var": "context"}]}} Its account uses the primary demand u and
local operating condition context. It does not give a primary-response law for
the auxiliary installation. Each trial starts the primary specimen at the same
unloaded mechanical or job condition. The auxiliary installation is not reset.
```

```
The recorded primary marker is the selected u, not a second measurement of the
response. This marker is especially distinct from the detector channel and
from the supervisory relay.

Controls have these meanings: u in [0, 1] sets the primary demand. For
context, 0 copies the primary request polarity; 1 reverses that polarity
before supervision. source in [0,1,2,3] means submit the maintenance-note
code; zero is the idle auxiliary setting. a=0 means disconnect the supervisor
service-message subscription for left; a=1 reconnects it. b does the same
independently for right. These controls do not remove a primary specimen or
edit its meter. A separate service receipt records a*source at left and
b*source at right: isolation prevents the selected service command from
reaching that specimen. A production RUN must deliver BOTH the primary target
and the requested auxiliary service code at both specimens. Calibration with
isolated sleeves is permitted and still returns ordinary measured responses.
```

Private worked solution. The following observations come from the released EXPAND reference policy. The fitted source effect at $s = (0, 1, 2, 3)$ is $(0, 1, 0, 0)$. The model and policy use Equations 8 and 9. Table 9 lists every reference action and raw measurement.

Table 9: C6-R1 Natural EXPAND reference. Outputs use the public units above. The final stopping action has no measurement.

Turn	Action	(u, c, s, a, b)	Raw outputs	Correction
1	TRIAL	$(1, 0, 0, 1, 1)$	$(0, 0)$	n/a
2	TRIAL	$(1, 0, 1, 1, 1)$	$(1, 1)$	n/a
3	TRIAL	$(1, 0, 2, 1, 1)$	$(0, 0)$	n/a
4	TRIAL	$(1, 0, 3, 1, 1)$	$(0, 0)$	n/a
5	TRIAL	$(1, 0, 0, 0, 1)$	$(0, 0)$	n/a
6	TRIAL	$(1, 0, 1, 0, 1)$	$(0, 1)$	n/a
7	TRIAL	$(1, 0, 0, 1, 0)$	$(0, 0)$	n/a
8	TRIAL	$(1, 0, 1, 1, 0)$	$(1, 0)$	n/a
9	RUN	$(0, 1, 1, 1, 1)$	$(1, 1)$	$(0, 0)$
10	STOP	n/a	n/a	n/a

At the previously untested production input, the committed model predicts $(1, 1)$ and the selected corrections deliver $(1, 1)$. The service receipt pair is $(1, 1)$. The reference completes source, selectivity, and blocking comparisons before this operation.

C.3 WORLD VARIANTS: THE SAME PROBLEM, DIFFERENT EXPLANATIONS

For the C5-R1 problem above, Table 10 compares raw volumes under the same four interventions. This is an explanatory replay, not a model trajectory. The manual predicts one microlitre in each channel at $u = 1, c = 0$.

Table 10: C5-R1 world comparison. Settings are (u, c, s, a, b) and each output pair is measured in microlitres.

Setting	RETAIN	REVISE	EXPAND
$(1, 0, 0, 1, 1)$	$(1, 1)$	$(1.5, 1.5)$	$(1.5, 1.5)$
$(1, 0, 1, 1, 1)$	$(1, 1)$	$(1.5, 1.5)$	$(2, 2)$
$(1, 0, 1, 0, 1)$	$(1, 1)$	$(1.5, 1.5)$	$(1.5, 2)$
$(1, 0, 1, 1, 0)$	$(1, 1)$	$(1.5, 1.5)$	$(2, 1.5)$

The first observation contradicts the manual in both REVISE and EXPAND. It does not distinguish a local correction from an auxiliary effect. The second observation supplies that distinction only in EXPAND. The two isolation measurements then test the individual paths.

C.4 Matched Narrative Skins

Table 11 prints the three opening narratives for the same C5-R1 instance. The source-effect values, controls, response units, manual, names, and budgets are unchanged. A Counter Prior statement is explicitly fallible. It is evidence for the agent to assess, not an extra constraint imposed on the simulator or on the countermodel language.

Table 11: Three renderings of C5-R1. The remaining prompt and physical response law are unchanged within each world variant.

Skin	Exact opening narrative
Neutral	Two dosing channels have independent collectors. A wash manifold accepts four bolus settings. Each dosing channel has an independently closable service-fluid branch.
Natural	Two synthesis channels deliver a requested dose into separate graduated collectors. The laboratory wash manifold can receive a separate bolus command. Service-fluid valves are independently closable during calibration. Primed trim reservoirs add or withdraw a quantity chosen before collection. Withdrawal cannot remove fluid that is not present.
Counter Prior	A fluidic isolation certificate states that wash boluses cannot enter either product collector. There are two independently valved service branches and two collection meters. The certificate is a testable assertion. A trim reservoir is commanded before the new collected volume is measured.

D Evaluation Details

This section distinguishes a response being well formed, an operation succeeding physically, a program being correct over its declared domain, and observed evidence supporting a causal claim.

D.1 Model and Action Space

A model predicts both raw outputs, and a policy specifies their corrections. Each is a pair of expressions in the public controls (u, c, s, a, b). The language permits rational constants, arithmetic, comparisons, Boolean operations, conditionals, minimum, and maximum. Each program is bounded to 512 expression nodes and depth 20, with constant and intermediate numerators and denominators bounded in magnitude by 10^{12}. Programs must be defined on every legal input and respect the output domain. Corrections must also lie within the physical capacity. These limits do not impose the generator's separable structure or the reference policy's prior.

Programs cannot access an outcome before it is observed, hidden state, elapsed time, files, or another episode. The manual and zero correction are used initially. Accepted updates are committed before the next action's outcome. Table 12 states the available actions and their observable effects. Appendix C.1 gives the action and expression syntax shown to agents.

Table 12 summarizes the action effects. Model-authored reasons are retained for qualitative inspection and are not scored. An invalid program update leaves the previously installed program in place. Scientific claims are evaluated from executable commitments and measurements, never by matching an explanation's wording to an answer key.

D.2 Production at Untested Inputs

A qualifying `RUN` uses an exact model and a policy valid over all operational inputs with both gates connected. The selected corrections must equal the committed policy's outputs, and both raw predictions and physical production must succeed. Programs are committed before the outcome. The complete input tuple must be unmeasured, not each individual value. Changing only the operation or correction does not make a setting new. Accepted updates in the same response take effect before execution, and later updates cannot change the evaluation of an earlier operation.

Model checks enumerate all 64 legal input tuples in binary families and 96 in the others. Policy checks use the 16 or 24 operational tuples with both gates connected. Physical success depends on

Table 12: Public action contract. Each submitted action consumes one of the 20 turns. Program updates are processed before the action.

Action	Observable effect
`TRIAL`	Sets u, c, s, a, b and measures both raw outputs, the primary marker, and the two service receipts. It does not apply production corrections.
`RUN`	Makes the same measurements and applies the selected corrections. Feedback reports delivered outputs, targets, receipts, and physical success.
`INSPECT`	Reads the stored meter and remaining budget without a new intervention. It cannot create a new source or path comparison.
`STOP`	Terminates the conversation without a measurement.
Invalid	Consumes a turn and returns an error, without supplying a valid intervention measurement.

the delivered endpoints and service receipts, not evidence flags. The contract requires $as = s$ and $bs = s$ in addition to corrected targets. Previously untested inputs may use $s = 0$.

D.3 Intervention Evidence

Basic Success requires a qualifying operation as defined above. In RETAIN and REVISE, *Discovery Success* additionally requires two valid measurements at fixed u, c, a, b, with $a = b = 1$, distinct source values, and identical raw output pairs. Either source value may be zero. This establishes invariance at the tested settings, not absence of an effect over the full domain.

In EXPAND, two diagnostic metrics describe partial evidence. *Source Test* requires a measured output contrast when only s changes (Equation 2). *Path Tests* requires certificates for both paths (Equation 3). Each combines a source contrast and a selective gate effect sharing a connected measurement. Both diagnostics use the complete episode record, including measurements after a qualifying `RUN`. Certificates can share measurements, use different primary settings, and occur in any order. Neither source value is required to be zero, and the fixed opposite gate may be connected or isolated. A certificate establishes a source contrast and a selective gate effect, not that the source contrast vanishes under isolation.

The evaluator searches all measured combinations, not a prescribed sequence. Each path may use its own source values and primary conditions. A gate contrast at a third source value is also accepted whenever it supplies the same evidence: if $y_j(1, s_0) \neq y_j(1, s_1)$, then $y_j(1, s_2)$ differs from at least one of them. A selective gate contrast at s_2 can therefore be paired with that connected measurement to form a three-setting witness. This argument holds at fixed primary conditions and opposite gate.

Only evidence collected up to and including the qualifying `RUN` counts toward that operation's *Discovery Success*. The `RUN` itself can supply a missing measured comparison, provided the model and policy were already committed and all other criteria hold. Textual assertions about an unperformed experiment do not count. An incomplete certificate is reported as incomplete evidence for the specified causal claim.

In EXPAND, *Discovery Success* adds a measurement at the other isolated source value, requiring $y_{\mathrm{L}}(0, s_0) = y_{\mathrm{L}}(0, s_1)$ for the left path and the analogous equality for the right path. Isolation must leave the other output unchanged at both source values. Both paths must satisfy these requirements by the same qualifying `RUN`. Evidence alone, without *Basic Success*, does not establish *Discovery Success*.

The criterion is applied uniformly to all 7 model configurations. Each certificate records the source and gate comparison turns and the qualifying operation.

D.4 Models and Inference Settings

The evaluation includes four locally hosted open-weight models and three frontier models accessed through provider APIs. All use the same text-only interaction protocol. Models receive the pub-

lic instructions and observations, then select actions and submit programs through the benchmark interface. The model descriptions below refer to the evaluated checkpoints and API identifiers.

- **Qwen3.5-4B and Qwen3.5-9B** are post-trained vision-language models whose language components have 4 and 9 billion parameters, respectively (Qwen Team, 2026a;b). Both combine Gated DeltaNet and attention layers and support native thinking. We use `Qwen/Qwen3.5-4B` and `Qwen/Qwen3.5-9B` in BF16 with thinking enabled. They provide two model sizes within the same family.
- **GPT-OSS-20B** is OpenAI's open-weight reasoning model with approximately 21 billion total parameters and 3.6 billion active parameters (OpenAI, 2026c). We run `openai/gpt-oss-20b` locally with its native MXFP4 weights and medium reasoning effort. It provides a non-Qwen open-weight baseline with adjustable reasoning.
- **Qwen3.8-27B** is a dense vision-language model with a 27-billion-parameter language component and configurable reasoning effort (Qwen Team, 2026c). We use the official `Qwen/Qwen3.8-27B-FP8` checkpoint at medium effort. This is a different model generation from the two smaller Qwen baselines.
- **Gemini 3.8 Flash** is a Google frontier model (Google DeepMind, 2026). We access `gemini-3.8-flash` through the Gemini API with medium thinking. Its multimodal inputs are not used in this text-only benchmark.
- **GPT-5.6-sol** is an OpenAI model with configurable reasoning effort (OpenAI, 2026a). We use the `gpt-5.6-sol` identifier through the Responses API at medium effort.
- **GPT-6 Astra** is an OpenAI model (OpenAI, 2026b). We use `gpt-6-astra` through the same Responses API and at the same medium effort setting as GPT-5.6-sol.

Table 13 summarizes the main configurations. Each primary response and permitted formatting repair has a 4,096-token output cap, including reasoning tokens where reported. Independent conversations retain all visible messages. External search and code execution are unavailable.

Table 13: Main evaluation settings. Context is the local total token capacity. Frontier models retain all visible messages within provider limits. Default means sampling parameters are not overridden.

Model	Weights	Reasoning	Decoding	Context	Runs
Qwen3.5-4B	BF16	Native on	Sampling	65,536	3
Qwen3.5-9B	BF16	Native on	Sampling	65,536	3
GPT-OSS-20B	MXFP4	Medium	Sampling	65,536	3
Qwen3.8-27B	FP8	Medium	Sampling	65,536	3
Gemini 3.8 Flash	Provider	Medium	$T = 0$	Provider	3
GPT-5.6-sol	Provider	Medium	Default	Provider	3
GPT-6 Astra	Provider	Medium	Default	Provider	3

All local models use temperature 1.0 and seeds 0, 1, and 2. Qwen uses top-p 0.95 and top-k 20, with presence penalty 1.5 for 4B and 9B and zero for 27B. GPT-OSS uses top-p 1.0, no top-k filter, and zero presence penalty. Frequency penalties are zero and repetition penalties are 1.0. Qwen3.5 supports native thinking but no medium effort parameter. Gemini uses temperature zero. GPT-5.6-sol and Astra leave temperature, top-p, and seed at provider defaults. A shared effort label does not equalize computation, and temperature zero does not guarantee identical provider responses.

Only visible final responses enter evaluation and conversation context. An invalid response permits one formatting repair with no new observation or scientific guidance. If the repaired response remains invalid, the action consumes a turn without a measurement. Responses are not automatically retried at a higher cap.

D.5 Solvability Reference

The reference demonstrates that exact modeling, measured causal evidence, and successful operation fit within one episode under a declared structural prior. Its model-fitting measurements and path comparisons serve distinct roles.

D.5.1 Fitted Model and Correction Policy

The reference assumes Equation 5 with stationary δ and h, independent path gates, and $h(0) = 0$. It reads the public manual and output operation but not the private values of δ or h, the mechanism variant, or the world label. This structural prior is not supplied to the language-model agents. The output group is rational addition, binary XOR, or addition modulo three. We write $x \ominus y = x \oplus (-_{\oplus} y)$ for its difference operation, where $-_{\oplus} y$ denotes the group inverse. In the XOR case, each element is its own inverse.

At $u = 1, c = 0, a = b = 1$, the reference measures $s = 0, 1, 2, 3$. Writing these observations as $\bar{y}_j(s)$, it estimates the local correction and source response by

$$\widehat{\delta}_j = \bar{y}_j(0) \ominus m_j(1, 0), \qquad \widehat{h}_j(s) = \bar{y}_j(s) \ominus \bar{y}_j(0). \tag{7}$$

The reference applies this fit separately to each unit. It represents the four estimated source effects with three nested conditionals over s, not by selecting a formula from the private generator. Let

$$\widehat{r}_j(q) = \widehat{\delta}_j \oplus g_j \widehat{h}_j(s). \tag{8}$$

The submitted programs are

$$\widehat{F}_j(q) = m_j(u, c) \oplus \widehat{r}_j(q), \qquad \pi_j(q) = -_{\oplus} \widehat{r}_j(q). \tag{9}$$

Thus the correction is the negative residual for additive outputs, the residual itself for XOR, and the negative residual modulo three for channel outputs. The construction uses only public control variables and fitted constants. The released mechanisms keep these programs within the expression limits and physical correction capacity in Appendix D.1.

D.5.2 Adaptive Measurement Schedule

One shared rule determines the measurements from the public responses. No step selects an action by a private world label.

1. Measure all four source settings at $u = 1, c = 0, a = b = 1$.
2. If a measured output pair differs from the idle pair, choose the smallest such source value s^*. At the same primary controls, measure both $s = 0$ and $s = s^*$ with $(a, b) = (0, 1)$, then with $(a, b) = (1, 0)$. Together with the earlier connected measurements, these complete both four-setting path certificates. If all pairs agree, omit these four trials.
3. Construct the programs in Equation 9. Submit both in the same response as a `RUN` at $u = 0, c = 1, a = b = 1$, with source s^* if one was found and $s = 1$ otherwise. Set the explicit corrections to $\pi(q)$ before receiving the outcome.
4. If the committed model predicts both raw outputs and production succeeds, submit `STOP`.

The trials that estimate the response function and those that establish path evidence have different roles. Under the structural prior, the first four measurements suffice to fit the response. In Expand, the isolation measurements test selective gate effects and whether the source contrast vanishes under isolation. These comparisons are collected even though the prior already determines their predictions.

D.5.3 Correctness and Action Budget

With both gates connected, Equation 5 gives $\bar{y}_j(s) = m_j(1, 0) \oplus \delta \oplus h(s)$. Since $h(0) = 0$, group cancellation in Equation 7 yields $\widehat{\delta}_j = \delta$ and $\widehat{h}_j(s) = h(s)$ at every source value. The source domain contains exactly those four values. By the assumed response structure, Equation 9 therefore gives $\widehat{F}(q) = F(q)$ over the full legal intervention domain. Its correction satisfies

$$F_j(q) \oplus \pi_j(q) = m_j(u, c) \oplus \widehat{r}_j(q) \oplus \big(-_{\oplus} \widehat{r}_j(q)\big) = m_j(u, c).$$

In particular, the policy satisfies the production requirements at every operational setting with $a = b = 1$. For the released additive mechanisms, the residual magnitude is at most 19, below the correction capacity of 32. Fluid withdrawal leaves the nonnegative target volume. XOR and channel corrections remain in their declared finite domains. The constructed programs use at most 56 expression nodes and depth nine, within the respective limits of 512 and 20.

Table 14: Full prompt-comparison results on 24 matched Neutral EXPAND episodes. Rates (%) are mean ± sample standard deviation over three runs.

Model	Prompt	Basic Success	Source Test	Path Tests	Discovery Success
Gemini 3.8 Flash	Direct	95.8 ± 4.2	79.2 ± 7.2	1.4 ± 2.4	0.0 ± 0.0
	General review	97.2 ± 4.8	87.5 ± 7.2	6.9 ± 6.4	0.0 ± 0.0
	Goal clarification	93.1 ± 6.4	83.3 ± 4.2	29.2 ± 0.0	0.0 ± 0.0
	All-claims clarification	97.2 ± 4.8	93.1 ± 2.4	41.7 ± 7.2	5.6 ± 4.8
	Experimental planning	97.2 ± 4.8	84.7 ± 2.4	8.3 ± 4.2	0.0 ± 0.0
GPT-5.6-sol	Direct	95.8 ± 4.2	52.8 ± 9.6	18.1 ± 9.6	0.0 ± 0.0
	General review	97.2 ± 2.4	45.8 ± 4.2	15.3 ± 6.4	0.0 ± 0.0
	Goal clarification	97.2 ± 2.4	91.7 ± 4.2	83.3 ± 7.2	0.0 ± 0.0
	All-claims clarification	100.0 ± 0.0	88.9 ± 6.4	81.9 ± 8.7	1.4 ± 2.4
	Experimental planning	97.2 ± 2.4	97.2 ± 2.4	76.4 ± 8.7	2.8 ± 4.8
GPT-6 Astra	Direct	100.0 ± 0.0	83.3 ± 4.2	27.8 ± 2.4	0.0 ± 0.0
	General review	100.0 ± 0.0	81.9 ± 13.4	18.1 ± 4.8	0.0 ± 0.0
	Goal clarification	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	11.1 ± 2.4
	All-claims clarification	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	98.6 ± 2.4
	Experimental planning	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	22.2 ± 9.6

In RETAIN and REVISE, $h \equiv 0$. Source variation cannot produce an auxiliary-effect witness or a path certificate. Instead, the four connected measurements have identical raw outputs, satisfying the invariant-comparison criterion. In EXPAND, nonconstant h with $h(0) = 0$ guarantees a source s^* with $h(s^*) \neq 0$. The connected measurements at 0 and s^* then differ, establishing the witness in Equation 2. For each unit, changing its gate from one to zero changes its source contrast from $h(s^*)$ to zero, while leaving the other unit's response unchanged at each source value. The four additional measurements thus complete both four-setting path certificates in Appendix D.3.

All calibration measurements have $u = 1, c = 0$. The final `RUN` uses $u = 0, c = 1$, so its complete input tuple is new. Both settings are legal in every released family. The model, policy, and explicit corrections are committed before this outcome, with all required evidence already in the record. Exact prediction and correction produce both primary targets, and $a = b = 1$ delivers the requested service code to both units. This operation therefore satisfies *Basic Success* and *Discovery Success* in every world. There are four initial trials, four further trials precisely when $h \not\equiv 0$, one `RUN`, and one `STOP`. The resulting counts are six actions for RETAIN and REVISE and ten for EXPAND.

Replay verifies this construction on all 216 released episodes. The argument depends on the declared structural prior, deterministic measurements, and the stated input domains and correction capacities.

D.6 PROMPT CONDITIONS

The comparison uses all 24 Neutral EXPAND cells for Gemini 3.8 Flash, GPT-5.6-sol, and GPT-6 Astra, with three runs per condition. Direct reuses the matched subsets of the main evaluation. Only the initial prompt changes under the four additional conditions. Model settings, response and repair caps, the 20-action budget, and the public problem remain fixed. Each episode uses a fresh conversation. Table 14 reports all four metrics. The prompt additions differ in length, and equal budget caps do not imply equal amounts of computation.

Experimental planning supplies a general comparison strategy without listing causal claims. Goal clarification specifies source and selectivity goals while clarifying control semantics. Neither addition explicitly requests blocking verification, so verification of that prediction requires agents to identify it in their own model. All-claims clarification adds the blocking goal to Goal clarification and assesses verification after the target claim has been supplied. The comparison measures evidence completion without assigning incomplete records to a specific internal stage.

The complete General review addition is:

> General task review: Before stopping, review whether your submitted model and policy satisfy the task requirements. Use the remaining interaction budget to verify and complete the task when necessary. Keep the original response format.

The complete Experimental planning addition is:

> General experimental planning guidance: Before selecting an experiment, identify which explanations it could distinguish. For each comparison, vary the factor you are testing while holding other relevant conditions fixed. Keep track of which claims are supported by matched observations and which remain uncertain. Before stopping, check whether your causal claims have experimental support. If a relevant comparison is missing, use the remaining action budget to obtain it. A discrepancy need not imply an additional causal influence. Keep the original response format.

The complete Goal clarification addition is:

> Task goals and assumptions: A correct response model and successful production do not by themselves establish experimental support for the causal claims in that model. For an explanation that extends the manual's causal scope, the measured record must support an effect of the auxiliary input on the raw responses and a selective effect of each connection control on its corresponding raw output. Selectivity means an effect on that output without an accompanying change in the other output at the conditions supported by the observations. Separate operation of the connection controls and their service receipts describe command delivery. They do not guarantee that a raw output is independent of the opposite connection control. Such independence is an additional assumption unless supported by the measured record. The evidence for these claims must be available no later than the production RUN that validates the committed model and policy at a previously untested setting. An auxiliary effect is not presumed to exist. Retaining the manual or correcting a relation within its scope remains appropriate when supported by the evidence. Keep the original response format.

All-claims clarification uses the complete Goal clarification addition above with only the following sentence inserted immediately before "The evidence for these claims must be available".

> The model's claim that isolation removes the auxiliary input's effect on each output also requires experimental support.

No intervention settings, comparison counts, or action sequence are added. The Direct-to-Goal comparison changes evidence targets and the interpretation of control descriptions together. In particular, the addition distinguishes command delivery from independence of the opposite output, not from blocking on the controlled path. The Goal-to-All-claims comparison retains this text and adds only the blocking target. Neither comparison establishes an agent's internal reason for accepting or leaving a prediction untested.

D.7 Evidence Completion and Operation Timing

Table 15 compares three evidence requirements on the same 1,080 stored episodes underlying the frontier prompt comparison. The 24 Neutral EXPAND episodes are repeated three times in each model–condition combination. For each requirement, we inspect all observations up to and including every `RUN` meeting *Basic Success*, and then the complete episode record. Rechecking these measurements reproduces all reported prompt-comparison counts.

The paired counts distinguish evidence available by a qualifying operation from evidence obtained at any time. The three failure columns partition *Discovery Success* failures. An episode without *Basic Success* is assigned to *Basic failed*. Among the others, *Missing* means that the episode ends without both four-setting path comparisons, while *Late* means that they are complete but no qualifying operation follows. There are no late-evidence failures of *Discovery Success* in this panel. The endpoint counts still distinguish incomplete evidence from failures of modeling or operation. In particular, Astra's Goal clarification condition has exact models and successful operation in every episode, but only 8 of 72 complete blocking evidence by episode end.

Alternative comparisons in All-claims episodes. We inspect the 58 Sol and 26 Gemini All-claims episodes that achieve *Basic Success* and both selective-effect certificates by a qualifying `RUN`, but end without both four-setting certificates. For each output j, let g_j denote its gate. Holding u, c and the opposite gate fixed, we search the measurements for the baseline-return equality $y_j(0, s) = y_j(1, 0)$ with $s > 0$. We separately check for source comparisons with $g_j = 0$ held fixed and for complete four-setting certificates. Table 16 reports coverage without changing the evidence criteria or prompt-comparison scores.

Table 15: Evidence completion in the prompt comparison, with 72 episodes per row. Each paired entry is the count by a qualifying `RUN` / at episode end. Source tests dependence, Selective requires both three-setting path certificates, and Blocking requires both four-setting path certificates. The first Blocking count equals the number of *Discovery Success* episodes. Missing, Late, and Basic failed partition the remaining episodes.

Model	Prompt	Source	Selective	Blocking	Missing	Late	Basic failed
Gemini 3.8 Flash	Direct	55/57	0/1	0/0	69	0	3
	General review	62/63	5/5	0/0	70	0	2
	Goal clarification	56/60	20/21	0/0	67	0	5
	All-claims clarification	65/67	30/30	4/4	66	0	2
	Experimental planning	59/61	6/6	0/0	70	0	2
GPT-5.6-sol	Direct	37/38	12/13	0/0	69	0	3
	General review	33/33	11/11	0/0	70	0	2
	Goal clarification	64/66	58/60	0/0	70	0	2
	All-claims clarification	64/64	59/59	1/1	71	0	0
	Experimental planning	68/70	53/55	2/2	68	0	2
GPT-6 Astra	Direct	60/60	20/20	0/0	72	0	0
	General review	59/59	13/13	0/0	72	0	0
	Goal clarification	72/72	72/72	8/8	64	0	0
	All-claims clarification	72/72	72/72	71/71	1	0	0
	Experimental planning	72/72	72/72	16/16	56	0	0

Table 16: Recorded comparisons in blocking-incomplete All-claims episodes. Counts pool three runs. Baseline return compares an isolated active output with its connected idle baseline at fixed primary inputs and opposite gate. The comparison types can overlap.

Comparison coverage	Sol	Gemini
Selected episodes	58	26
Baseline return on both paths	53	13
Baseline return on one path	0	1
Baseline return on neither path	5	12
Isolated source comparison on both paths	7	10
Isolated source comparison on one path	2	4
Isolated source comparison on neither path	49	12
Four-setting certificate on one path	1	3
Four-setting certificates on both paths	0	0

For every path with a baseline-return pair, at least one such pair also shares its active isolated setting with an observed selective gate comparison and an output-changing connected source comparison. However, inferring $y_j(0, s) = y_j(0, 0)$ from baseline return requires the further equality $y_j(0, 0) = y_j(1, 0)$. This is an additional premise about raw outputs. Zero service-command delivery at an idle source does not guarantee that changing a gate leaves its output unchanged. None of Sol's baseline-return pairs includes this measured idle gate comparison. One Gemini episode includes it, on one path. Without it, baseline return gives conditional support for blocking rather than the source-under-isolation comparison required by Equation 4. Conversely, the 7 Sol and 10 Gemini episodes with isolated source comparisons on both paths do not combine them with output-changing connected comparisons at matching primary inputs, opposite-gate settings, and source values into both four-setting certificates. Some measured four-setting grids also use source values with no connected output contrast. Thus incomplete evidence includes missing isolated source comparisons and failure to combine them with a corresponding connected source contrast.

The final `STOP` explanations in these 84 episodes do not explicitly cite the baseline-return equality as their justification for blocking. Earlier explanations can make the connection more concrete. In Sol's C3-R1 episode (run 1), the pre-production explanation cites "restoration of each idle-service baseline under isolation." At $u = c = 0$, turns 1–4 measure (s, a, b) equal to $(0, 1, 1)$, $(1, 1, 1)$, $(1, 0, 1)$, and $(1, 1, 0)$, respectively. The raw outputs are $(1, 1)$, $(0, 0)$, $(1, 0)$, and $(0, 1)$. Each isolated output returns to its connected idle baseline, but neither path is measured at two source values under fixed isolation. Gemini's C2-R2 episode (run 2) similarly describes a both-gates-isolated trial at $s = 2$ as

a test that isolation removes the effect. It never changes the source with either gate held isolated. These examples connect the stated verification goals to the comparisons actually performed.

E ADDITIONAL RESULTS

Table 17 reports the applicable metrics on the same runs as Figure 2. Each run evaluates the same 216 episodes, with 72 per world across all families and skins. All entries are means and sample standard deviations over three complete runs and use the 4,096-token response cap. Output-budget comparisons and prompt controls are separate conditions and are not included. Repeated runs, world variants, and skins are not independent mechanisms. Run standard deviations describe repeated execution on this fixed release, not uncertainty about performance on unseen mechanisms. Our primary metric is *Discovery Success*. *Basic Success* requires correct modeling and successful operation in every world. Discovery additionally requires invariant comparisons at tested settings in RETAIN and REVISE, and source, selectivity, and blocking evidence in EXPAND (Section 4.2).

Table 17: Full benchmark results (%). All metrics within a row use the same episodes and runs. R is the number of runs, each with 72 episodes per world. For $R = 3$, entries show mean $\pm$ sample standard deviation. Source and Path Tests apply only to EXPAND. A dash denotes not applicable.

Model	R	Theory Accuracy	Control Success	Basic Success	Source Test	Path Tests	Discovery Success
RETAIN							
Qwen3.5-4B	3	97.7 ± 1.6	91.2 ± 7.1	86.1 ± 10.5	—	—	35.2 ± 9.9
Qwen3.5-9B	3	99.5 ± 0.8	97.2 ± 1.4	94.0 ± 4.5	—	—	65.7 ± 8.5
GPT-OSS-20B	3	98.1 ± 0.8	97.2 ± 1.4	82.9 ± 2.1	—	—	7.9 ± 0.8
Qwen3.8-27B	3	99.5 ± 0.8	100.0 ± 0.0	95.4 ± 1.6	—	—	38.0 ± 9.2
Gemini 3.8 Flash	3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	—	—	19.9 ± 2.1
GPT-5.6-sol	3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	—	—	49.5 ± 2.9
GPT-6 Astra	3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	—	—	94.0 ± 2.1
REVISE							
Qwen3.5-4B	3	31.9 ± 13.7	50.0 ± 6.4	9.7 ± 2.4	—	—	3.7 ± 0.8
Qwen3.5-9B	3	60.6 ± 2.1	60.2 ± 4.5	38.4 ± 2.1	—	—	25.9 ± 4.2
GPT-OSS-20B	3	74.5 ± 6.3	69.4 ± 2.4	53.2 ± 2.9	—	—	8.3 ± 2.4
Qwen3.8-27B	3	100.0 ± 0.0	99.1 ± 0.8	99.1 ± 0.8	—	—	54.6 ± 3.2
Gemini 3.8 Flash	3	99.5 ± 0.8	99.5 ± 0.8	99.5 ± 0.8	—	—	26.9 ± 1.6
GPT-5.6-sol	3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	—	—	33.3 ± 1.4
GPT-6 Astra	3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	—	—	36.1 ± 1.4
EXPAND							
Qwen3.5-4B	3	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	35.2 ± 5.3	1.9 ± 1.6	0.0 ± 0.0
Qwen3.5-9B	3	1.9 ± 0.8	1.9 ± 0.8	0.0 ± 0.0	64.8 ± 0.8	5.1 ± 0.8	0.0 ± 0.0
GPT-OSS-20B	3	0.0 ± 0.0	2.8 ± 1.4	0.0 ± 0.0	19.0 ± 2.1	0.0 ± 0.0	0.0 ± 0.0
Qwen3.8-27B	3	41.7 ± 6.4	47.7 ± 6.6	40.3 ± 6.4	54.2 ± 6.4	0.9 ± 0.8	0.0 ± 0.0
Gemini 3.8 Flash	3	96.8 ± 2.9	95.8 ± 3.7	95.8 ± 3.7	83.3 ± 3.7	2.3 ± 0.8	0.0 ± 0.0
GPT-5.6-sol	3	98.1 ± 2.1	98.1 ± 2.1	98.1 ± 2.1	49.1 ± 1.6	21.3 ± 2.9	0.0 ± 0.0
GPT-6 Astra	3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	81.0 ± 5.6	31.0 ± 1.6	0.5 ± 0.8

Metrics follow Section 4.2 and Appendix D.3. *Source Test* and *Path Tests* use the complete recorded episode. *Discovery Success* requires its evidence by the qualifying operation, including that operation's measurement. The countermodel analysis uses the EXPAND *Basic Success* subset of these same runs, examining each episode separately without pooling its measurements with another run. All three runs contribute equally.

E.1 Local Model Prompt Comparison

Table 18 extends the prompt comparison to Qwen3.5-4B, Qwen3.5-9B, and GPT-OSS-20B. Each condition uses the same 24 Neutral EXPAND episodes and three sampling seeds. The prompt additions are identical to those in Appendix D.6. Direct reuses the matching episodes from the main evaluation. All conditions retain the 4,096-token response cap, 65,536-token context, model-specific sampling and reasoning settings, and 20-action limit. The prompted runs use a batch of 24, compared with 35 for Direct Qwen3.5-4B and 30 for Direct Qwen3.5-9B and GPT-OSS-20B. Seeds therefore identify repeated runs, not matched random draws across conditions.

Table 18: Local model prompt comparison on 24 Neutral EXPAND episodes. Rates (%) are mean ± sample standard deviation over three runs.

Model	Prompt	Basic Success	Source Test	Path Tests	Discovery Success
Qwen3.5-4B	Direct	0.0 ± 0.0	43.1 ± 12.7	4.2 ± 4.2	0.0 ± 0.0
	General review	0.0 ± 0.0	26.4 ± 2.4	1.4 ± 2.4	0.0 ± 0.0
	Goal clarification	0.0 ± 0.0	52.8 ± 6.4	12.5 ± 4.2	0.0 ± 0.0
	Experimental planning	0.0 ± 0.0	40.3 ± 10.5	1.4 ± 2.4	0.0 ± 0.0
Qwen3.5-9B	Direct	0.0 ± 0.0	62.5 ± 4.2	2.8 ± 4.8	0.0 ± 0.0
	General review	0.0 ± 0.0	72.2 ± 2.4	4.2 ± 4.2	0.0 ± 0.0
	Goal clarification	1.4 ± 2.4	83.3 ± 0.0	22.2 ± 10.5	0.0 ± 0.0
	Experimental planning	0.0 ± 0.0	69.4 ± 6.4	6.9 ± 6.4	0.0 ± 0.0
GPT-OSS-20B	Direct	0.0 ± 0.0	19.4 ± 6.4	0.0 ± 0.0	0.0 ± 0.0
	General review	0.0 ± 0.0	16.7 ± 4.2	0.0 ± 0.0	0.0 ± 0.0
	Goal clarification	0.0 ± 0.0	15.3 ± 2.4	0.0 ± 0.0	0.0 ± 0.0
	Experimental planning	0.0 ± 0.0	15.3 ± 2.4	0.0 ± 0.0	0.0 ± 0.0

Goal clarification improves evidence collection for both Qwen models. Relative to Direct, *Path Tests* increase from 4.2% to 12.5% for 4B and from 2.8% to 22.2% for 9B. However, *Basic Success* remains zero for 4B and reaches only 1.4% for 9B, corresponding to one successful model-based operation in 72 trajectories. GPT-OSS-20B achieves no *Basic Success* under any condition. *Discovery Success* is zero throughout. Under these settings, clearer goals help some local models collect evidence but rarely suffice to complete both modeling and experimentally supported operation.

E.2 Outcome Breakdown

Table 19: EXPAND outcome counts (trajectories), using the categories in Figure 3. Each row sums to $n = 216$ across three runs.

Model	n	Both incomplete	Basic Success failed	Evidence incomplete	Blocking incomplete	Discovery Success
Qwen3.5-4B	216	212	4	0	0	0
Qwen3.5-9B	216	205	11	0	0	0
GPT-OSS-20B	216	216	0	0	0	0
Qwen3.8-27B	216	127	2	87	0	0
Gemini 3.8 Flash	216	8	1	203	4	0
GPT-5.6-sol	216	3	1	167	45	0
GPT-6 Astra	216	0	0	149	66	1

Table 19 gives the counts underlying Figure 3. Both incomplete denotes failure of both *Basic Success* and evidence collection. Basic Success failed denotes evidence without successful model-based operation, and Evidence incomplete denotes *Basic Success* without complete source and selectivity evidence. Blocking incomplete denotes model-based use with source and selectivity evidence by a qualifying `RUN`, but without complete blocking evidence by a qualifying operation. *Discovery Success* requires all three evidence components by the same qualifying `RUN`. Classification uses each recorded episode, not subtraction of marginal success rates. Some 4B and 9B episodes collect

source and selectivity evidence without achieving *Basic Success*. Requiring the source contrast to vanish under isolation yields EXPAND *Discovery Success* of 0.5% for Astra and zero for the other models (Table 17).

E.3 FAMILY AND SKIN RESULTS

Each family contributes four mechanism variants and three skins per world. Skin comparisons fix all 24 family and mechanism variant combinations. Table 20 reports family-level results, and Table 21 compares the three skins.

Table 20: Basic Success (%) by family. Entries are mean ± sample standard deviation over three runs, each containing 12 episodes per family and world.

Model	Family	RETAIN	REVISE	EXPAND
Qwen3.5-4B	C1	77.8 ± 19.2	22.2 ± 12.7	0.0 ± 0.0
	C2	91.7 ± 0.0	2.8 ± 4.8	0.0 ± 0.0
	C3	83.3 ± 14.4	5.6 ± 4.8	0.0 ± 0.0
	C4	80.6 ± 26.8	0.0 ± 0.0	0.0 ± 0.0
	C5	88.9 ± 9.6	5.6 ± 4.8	0.0 ± 0.0
	C6	94.4 ± 4.8	22.2 ± 4.8	0.0 ± 0.0
Qwen3.5-9B	C1	94.4 ± 4.8	38.9 ± 19.2	0.0 ± 0.0
	C2	91.7 ± 8.3	44.4 ± 17.3	0.0 ± 0.0
	C3	91.7 ± 8.3	50.0 ± 8.3	0.0 ± 0.0
	C4	94.4 ± 9.6	41.7 ± 14.4	0.0 ± 0.0
	C5	97.2 ± 4.8	25.0 ± 8.3	0.0 ± 0.0
	C6	94.4 ± 4.8	30.6 ± 4.8	0.0 ± 0.0
GPT-OSS-20B	C1	83.3 ± 8.3	41.7 ± 8.3	0.0 ± 0.0
	C2	88.9 ± 4.8	80.6 ± 12.7	0.0 ± 0.0
	C3	83.3 ± 0.0	30.6 ± 9.6	0.0 ± 0.0
	C4	72.2 ± 4.8	52.8 ± 12.7	0.0 ± 0.0
	C5	83.3 ± 8.3	72.2 ± 12.7	0.0 ± 0.0
	C6	86.1 ± 4.8	41.7 ± 14.4	0.0 ± 0.0
Qwen3.8-27B	C1	97.2 ± 4.8	97.2 ± 4.8	27.8 ± 17.3
	C2	100.0 ± 0.0	100.0 ± 0.0	52.8 ± 12.7
	C3	94.4 ± 4.8	100.0 ± 0.0	27.8 ± 9.6
	C4	91.7 ± 8.3	97.2 ± 4.8	52.8 ± 12.7
	C5	94.4 ± 4.8	100.0 ± 0.0	69.4 ± 12.7
	C6	94.4 ± 9.6	100.0 ± 0.0	11.1 ± 4.8
Gemini 3.8 Flash	C1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	C2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	C3	100.0 ± 0.0	100.0 ± 0.0	91.7 ± 8.3
	C4	100.0 ± 0.0	100.0 ± 0.0	97.2 ± 4.8
	C5	100.0 ± 0.0	97.2 ± 4.8	94.4 ± 9.6
	C6	100.0 ± 0.0	100.0 ± 0.0	91.7 ± 8.3
GPT-5.6-sol	C1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	C2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	C3	100.0 ± 0.0	100.0 ± 0.0	94.4 ± 9.6
	C4	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	C5	100.0 ± 0.0	100.0 ± 0.0	97.2 ± 4.8
	C6	100.0 ± 0.0	100.0 ± 0.0	97.2 ± 4.8
GPT-6 Astra	C1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	C2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	C3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	C4	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	C5	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	C6	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0

For Astra in EXPAND, the skin difference appears in evidence collection, with *Basic Success* at 100% in every skin. Its *Path Tests* rates are 45.8% for Counter Prior and 27.8% for Neutral, while

Table 21: Basic Success (%) by skin. Entries are mean ± sample standard deviation over three runs. Each run covers the same 24 family and mechanism variant combinations per skin and world.

Model	World	Neutral	Natural	Counter Prior
Qwen3.5-4B	RETAIN	84.7 ± 12.7	91.7 ± 11.0	81.9 ± 12.0
	REVISE	8.3 ± 4.2	11.1 ± 6.4	9.7 ± 2.4
	EXPAND	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Qwen3.5-9B	RETAIN	95.8 ± 4.2	91.7 ± 8.3	94.4 ± 4.8
	REVISE	30.6 ± 8.7	48.6 ± 2.4	36.1 ± 6.4
	EXPAND	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
GPT-OSS-20B	RETAIN	79.2 ± 7.2	90.3 ± 6.4	79.2 ± 7.2
	REVISE	45.8 ± 0.0	56.9 ± 9.6	56.9 ± 14.6
	EXPAND	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Qwen3.8-27B	RETAIN	93.1 ± 2.4	95.8 ± 4.2	97.2 ± 2.4
	REVISE	98.6 ± 2.4	98.6 ± 2.4	100.0 ± 0.0
	EXPAND	44.4 ± 13.4	37.5 ± 4.2	38.9 ± 2.4
Gemini 3.8 Flash	RETAIN	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	REVISE	98.6 ± 2.4	100.0 ± 0.0	100.0 ± 0.0
	EXPAND	95.8 ± 4.2	97.2 ± 2.4	94.4 ± 6.4
GPT-5.6-sol	RETAIN	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	REVISE	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	EXPAND	95.8 ± 4.2	100.0 ± 0.0	98.6 ± 2.4
GPT-6 Astra	RETAIN	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	REVISE	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	EXPAND	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0

the corresponding *Source Test* rates are 77.8% and 83.3%. These are matched tasks, not matched random draws. Counter Prior's testable assertion of no auxiliary influence may make conflicting observations more salient and encourage scrutiny.

E.4 EXPLORATION WITHOUT BOUNDARY EXPANSION

The following analyses distinguish exploration from completing the invariant comparisons used to evaluate evidence in RETAIN and REVISE. All quantities are computed from the same recorded experiments.

Table 22: Auxiliary-input exploration in the main evaluation (%). Entries are mean ± sample standard deviation over three runs of 72 episodes per world. Matched Trials differ only in source with both paths connected. Output change is not required.

Model	RETAIN		REVISE	
	Active Trials	Matched Trials	Active Trials	Matched Trials
Qwen3.5-4B	69.4 ± 11.4	16.2 ± 8.4	66.7 ± 6.1	14.8 ± 7.6
Qwen3.5-9B	98.1 ± 0.8	44.4 ± 6.1	95.4 ± 3.5	55.6 ± 1.4
GPT-OSS-20B	22.7 ± 3.5	0.5 ± 0.8	22.7 ± 6.6	0.5 ± 0.8
Qwen3.8-27B	86.6 ± 4.0	35.2 ± 8.0	83.8 ± 5.8	44.9 ± 4.2
Gemini 3.8 Flash	100.0 ± 0.0	12.5 ± 1.4	99.5 ± 0.8	25.0 ± 3.7
GPT-5.6-sol	100.0 ± 0.0	49.5 ± 2.9	100.0 ± 0.0	32.9 ± 2.1
GPT-6 Astra	100.0 ± 0.0	93.1 ± 2.4	100.0 ± 0.0	32.4 ± 2.9

Only valid measured `TRIAL` actions count. Active trials use $s > 0$. Matched trials differ in source at fixed primary inputs with both gates connected. The source pair may include zero or two positive values. Every recorded matched pair in these RETAIN and REVISE episodes has identical outputs. These comparisons record invariance at tested settings, not absence of an effect throughout the domain. These exploration rates do not require successful operation or restrict comparisons to the prefix ending at a qualifying `RUN`.

Table 23 reports how often successful model-based operation is accompanied by an invariant comparison. *Discovery Success* counts *Basic Success* with two measurements at identical u, c, a, b, with $a = b = 1$, different source values, and unchanged raw outputs. Both measurements must be available by the same qualifying `RUN`, including its own measurement. Valid `TRIAL` and `RUN` measurements count here, whereas *Matched Trials* above counts only `TRIAL`s over the full episode. A later comparison cannot justify an earlier operation. All rates use the same 72 episodes per world per run, not only successful episodes. The comparison does not certify absence of a source effect outside the tested settings.

Table 23: Basic Success and Discovery Success in RETAIN and REVISE (%). Entries are mean ± sample standard deviation over three runs.

Model	RETAIN		REVISE	
	Basic Success	Discovery Success	Basic Success	Discovery Success
Qwen3.5-4B	86.1 ± 10.5	35.2 ± 9.9	9.7 ± 2.4	3.7 ± 0.8
Qwen3.5-9B	94.0 ± 4.5	65.7 ± 8.5	38.4 ± 2.1	25.9 ± 4.2
GPT-OSS-20B	82.9 ± 2.1	7.9 ± 0.8	53.2 ± 2.9	8.3 ± 2.4
Qwen3.8-27B	95.4 ± 1.6	38.0 ± 9.2	99.1 ± 0.8	54.6 ± 3.2
Gemini 3.8 Flash	100.0 ± 0.0	19.9 ± 2.1	99.5 ± 0.8	26.9 ± 1.6
GPT-5.6-sol	100.0 ± 0.0	49.5 ± 2.9	100.0 ± 0.0	33.3 ± 1.4
GPT-6 Astra	100.0 ± 0.0	94.0 ± 2.1	100.0 ± 0.0	36.1 ± 1.4

E.5 OUTPUT-BUDGET SENSITIVITY

We compare 4,096 and 16,384 primary output tokens on all 24 Neutral EXPAND episodes for Qwen3.5-4B, Qwen3.5-9B, and GPT-OSS-20B, using seeds 0, 1, and 2. Each model contributes 72 trajectories per condition. The lower-budget results are the same episode IDs and seeds from the main sampling evaluation. The higher-budget condition starts fresh conversations for every episode, without selecting failures or capped outputs. Model revisions, quantization, prompts, model-specific sampling and reasoning settings, the 65,536-token context capacity, and the 20-action limit remain unchanged. The primary allowance includes reasoning tokens. Formatting repair retains its 4,096-token allowance. No conversation history is trimmed.

Table 24: Output-budget comparison on 24 matched Neutral EXPAND episodes. Rates (%) are mean ± sample standard deviation over three seeds. The cap applies to each primary response.

Model	Cap (tokens)	Basic Success	Source Test	Path Tests	Discovery Success
Qwen3.5-4B	4,096	0.0 ± 0.0	43.1 ± 12.7	4.2 ± 4.2	0.0 ± 0.0
Qwen3.5-4B	16,384	0.0 ± 0.0	31.9 ± 6.4	1.4 ± 2.4	0.0 ± 0.0
Qwen3.5-9B	4,096	0.0 ± 0.0	62.5 ± 4.2	2.8 ± 4.8	0.0 ± 0.0
Qwen3.5-9B	16,384	1.4 ± 2.4	76.4 ± 10.5	4.2 ± 4.2	0.0 ± 0.0
GPT-OSS-20B	4,096	0.0 ± 0.0	19.4 ± 6.4	0.0 ± 0.0	0.0 ± 0.0
GPT-OSS-20B	16,384	0.0 ± 0.0	16.7 ± 8.3	0.0 ± 0.0	0.0 ± 0.0

Discovery Success remains zero for every model and seed at both caps (Table 24). Other outcomes change. Qwen3.5-9B achieves *Basic Success* in one higher-budget trajectory, which lacks complete path evidence. Neither 4B nor GPT-OSS-20B achieves *Basic Success* at either cap. Source and path evidence also vary between conditions.

All higher-budget trajectories pass replay and configuration checks. Task IDs and seed numbers are matched, but different batch compositions mean that the sampling draws need not be identical.

E.6 FRONTIER RESPONSE-LIMIT DIAGNOSTICS

We distinguish an episode containing an output-limited response from a failure to complete experimental evidence. A capped episode contains at least one primary response whose recorded termina-

tion reason indicates the output limit. Formatting-repair responses are counted separately. Table 25 pools all three worlds and runs in the main evaluation, giving 648 episodes per model. Gemini has 124 capped episodes, while Sol and Astra have none. Unparsed responses count individual responses that remain invalid after the permitted formatting repair, not episodes.

Table 25: Frontier response limits in the main evaluation. Capped episodes contain a primary response ending at the output limit. Unparsed responses remain invalid after formatting repair.

Model	Episodes	Capped episodes	Unparsed responses
Gemini 3.8 Flash	648	124	11
GPT-5.6-sol	648	0	11
GPT-6 Astra	648	0	0

Table 26 restricts the source and selectivity analysis to EXPAND episodes with *Basic Success* and no capped primary response. It retains the four-model cohort used in Section 5.2. Among Gemini's 101 retained episodes, 99 still lack source or selectivity evidence. All 167 Sol and 149 Astra failures in this analysis remain because none contains a capped primary response. Every retained evidence gap admits an executable alternative and ends with `STOP` while actions remain. These counts concern source and selectivity evidence, not the additional blocking requirement of *Discovery Success*.

Table 26: Evidence analysis without capped primary responses in the main EXPAND evaluation. Eligible episodes satisfy Basic Success. Ratios report uncapped/eligible, gaps/uncapped, alternatives/gaps, and early stops/gaps. Gaps concern source or selectivity evidence.

Model	Uncapped	Evidence gaps	Alternatives	Early stops
Qwen3.8-27B	16/87	16/16	16/16	16/16
Gemini 3.8 Flash	101/207	99/101	99/99	99/99
GPT-5.6-sol	212/212	167/212	167/167	167/167
GPT-6 Astra	216/216	149/216	149/149	149/149

For Gemini Goal clarification, Table 27 reports response-condition subsets of the 72 episodes in the matched prompt comparison. Excluding capped primary responses leaves 34 episodes, none with a capped repair. *All responses normal* further requires normal provider endings for every primary and repair response. *Clean + Basic Success* also excludes invalid responses and actions and requires *Basic Success*. Of its 8 episodes, 6 still lack source or selectivity evidence, and none achieves *Discovery Success*. All eight stop with unused action budget. Thus, incomplete evidence is also observed without capped or abnormally terminated responses. These response-based subsets do not estimate the effect of increasing the output budget, and the Goal clarification breakdown does not describe the other prompt conditions.

Table 27: Gemini Goal clarification by response condition. Counts pool three runs of 24 Neutral EXPAND episodes. Each row has its own denominator n. Clean additionally excludes invalid responses and actions. Source Test and Path Tests use the full episode, while Discovery Success requires complete evidence by a qualifying RUN.

Subset	n	Basic Success	Source Test	Path Tests	Discovery Success
All	72	67	60	21	0
No primary cap	34	32	24	6	0
No primary cap + Basic Success	32	32	22	6	0
All responses normal	10	9	8	2	0
Clean + Basic Success	8	8	6	2	0

F EVIDENCE ANALYSIS

F.1 EVIDENCE SUFFICIENCY THROUGH EXECUTABLE ALTERNATIVES

This analysis makes source and selectivity evidence gaps explicit by constructing executable alternatives that fit every recorded measurement while denying a causal claim in the submitted model. The alternatives provide witnesses to unresolved claims rather than a separate failure rate.

The analysis includes all 722 EXPAND episodes satisfying *Basic Success* across the three main runs. Of these, 116 complete source and selectivity evidence by a qualifying `RUN`, while 606 do not. Selection depends on model correctness and successful operation, not the evidence score. Table 28 reports eligible denominators and results. Alternatives use the bounded response language in Appendix D.1 and satisfy the assumptions stated in the task (Section 2).

Table 28: Episode-end evidence analysis of Basic Success episodes. (a) Percentages are three-run means ± sample standard deviations, using each run's eligible denominator. (b) Pooled counts. Paths by RUN counts source and selectivity certificates completed by a qualifying `RUN`, without assessing blocking. Unlike the full-episode *Path Tests* diagnostic, this count requires evidence by the operation. Alternatives are checked against the complete episode record.

(a) Rates (%) across three runs			
Model	Episodes (n)	Paths by RUN	Alternatives remain
Qwen3.8-27B	87	0.0 ± 0.0	100.0 ± 0.0
Gemini 3.8 Flash	207	1.9 ± 0.7	98.1 ± 0.7
GPT-5.6-sol	212	21.2 ± 3.8	78.8 ± 3.8
GPT-6 Astra	216	31.0 ± 1.6	69.0 ± 1.6
(b) Counts summed across three runs			
Qwen3.8-27B	87	0	87
Gemini 3.8 Flash	207	4	203
GPT-5.6-sol	212	45	167
GPT-6 Astra	216	67	149
Total	722	116	606

For each episode, let $\hat{F}$ be the model publicly committed before its first successful `RUN` at previously untested inputs using an exact model and a valid policy. We examine the observed evidence immediately before this operation, through its outcome, and at episode end. The construction receives only the public contract, this program, and the relevant measurements. It does not use private response equations, new simulator outcomes, or the agent's reasoning text. A countermodel h' must disagree about the causal property being tested, not merely use different syntax, and must satisfy

$$h'(q_i) = y_i \quad \text{for all } (q_i, y_i) \in D, \qquad h'(q^*) \neq \hat{F}(q^*) \quad \text{for some unmeasured legal } q^*. \tag{10}$$

The query q^* is a hypothetical `TRIAL` that distinguishes the two programs. It is not executed or added to the agent's evidence. If h' were the response law, every recorded measurement would remain unchanged, yet $\hat{F}$ would be wrong at q^*. This is a counterfactual consequence of the compatible alternative, not an observed error in the released world or a new agent rollout.

The alternatives replace only the raw-response program. Legal settings, output domains, the primary marker u, service receipts (as, bs), correction rules, and production targets remain fixed. We reconstruct the recorded markers, receipts, and corrected outcomes with each countermodel to check these explicit public rules. The receipt rule guarantees that a disconnected service line does not receive the selected command. We do not equate it with a guarantee that every possible source effect on the raw response disappears under isolation.

For source dependence, an observed contrast satisfying Equation 2 excludes every deterministic source-independent response law. Otherwise we construct a response table over $x = (u, c, a, b)$,

fitting every observed value and setting unobserved entries to zero. Its output is independent of s by construction. We compile the table into the public expression language and check the program on all legal settings.

For each path, we test a restricted family of alternatives that cannot produce both a source contrast and a selective effect of that path's gate under the required matched conditions. Let g_j be the tested gate and g_{-j} the other gate. At each fixed (u, c, g_{-j}), the tested output is either constant across s when $g_j = 1$, or invariant to g_j at every source setting. The other output is unchanged from $\hat{F}$. Such an alternative cannot exhibit the linked source and selective gate contrasts required by the three-setting certificate for that path. When a local path certificate is observed, it rules out every alternative in this family. Failed construction is reported as unknown.

Every saved countermodel is checked against all collected measurements and all legal inputs, including the causal claim it denies. All programs satisfy the shared 512-node and depth-20 limits. The largest uses 197 nodes across both outputs. No component check returns an unknown result. These are alternatives within the admissible response-law class.

Compatible alternatives remain in all 606 failures. The 116 episodes completing source and selectivity evidence exclude the tested families of alternative explanations. Countermodels must satisfy every observation and the program limits, so their existence is checked rather than inferred solely from a missing certificate.

F.2 Analyzing Blocking Claims

We extend the analysis to the model's prediction that isolation removes the source effect. The cohort remains all 722 EXPAND episodes with *Basic Success*. Of these, 115 complete source and selectivity evidence by a qualifying `RUN` but fail *Discovery Success*. We analyze their complete episode records, keeping the model committed before the first qualifying operation fixed.

For output j, fix the primary inputs and the opposite gate. Each pair of observed connected settings with unequal outputs defines a source contrast (s_0, s_1). The submitted model predicts equal outputs at the corresponding isolated settings. We search for a program h' that fits every observation, retains every prediction with $g_j = 1$, and instead satisfies

$$h'_j(0, s_0) \neq h'_j(0, s_1) \tag{11}$$

for every such observed connected contrast. The other output remains unchanged. We assign isolated responses consistent with these inequalities and the observations, then express the resulting function as a program defined for every legal input within the task's program limits. Binary and ternary output restrictions are enforced. If a connected contrast already has two equal measured isolated responses, that observed pair excludes this family. This is a claim-specific analysis tied to the source contrasts in the record, not a test of global identification at unmeasured primary settings.

Table 29 reports the 115 episodes that isolate the added blocking requirement. Every episode admits at least one blocking-denying program. The largest uses 203 nodes, below the 512-node limit. All programs match every recorded measurement, marker, receipt, and corrected outcome. Their programs are evaluated at every legal input, and each tested pair retains a nonzero isolated contrast. The one episode achieving *Discovery Success* supplies measured blocking on both paths and excludes these alternatives at its observed contrasts. Combining this analysis with the source and selectivity analysis covers all 721 *Basic Success* episodes failing *Discovery Success*, using the same trajectory hashes and committed programs. The two analyses provide explicit witnesses for different missing claims.

The construction preserves all predictions with $a = b = 1$. Consequently, it also preserves every operating outcome under the fixed policy. Its distinguishing experiment is an isolation test, not a production failure. This separates a model's successful use under connection from experimental support for its claim about removing the influence.

Blocking versus cancellation. A compact C2-R1 example illustrates why the isolated source comparison matters. Consider the released response law F and an additive alternative F',

$$F_{\mathrm{L}}(q) = 2u + c + 1 + 2as, \quad F'_{\mathrm{L}}(q) = 2u + c + 1 + 2s - 2(1 - a),$$
$$F_{\mathrm{R}}(q) = 2u + c + 1 + 2bs, \quad F'_{\mathrm{R}}(q) = F_{\mathrm{R}}(q).$$

Table 29: Blocking-denying alternatives among episodes completing source and selectivity evidence but failing Discovery Success. Counts pool three runs.

Model	Episodes	Alternatives remain
Gemini 3.8 Flash	4	4
GPT-5.6-sol	45	45
GPT-6 Astra	66	66
Total	115	115

The alternative attributes the left response to an ungated source effect and an independent -2 mm shift when the left gate is isolated. At $u = 1, c = 0, b = 1$, the two programs agree on both outputs at the three settings below that form a left-path certificate. Only the fourth setting distinguishes them.

a	s	$F(q)$	$F'(q)$	Role
1	0	$(3, 3)$	$(3, 3)$	Connected idle baseline
1	1	$(5, 5)$	$(5, 5)$	Connected source contrast
0	1	$(3, 5)$	$(3, 5)$	Selective gate contrast
0	0	$(3, 3)$	$(1, 3)$	Distinguishing isolated comparison

Returning to the connected idle baseline at $(a, s) = (0, 1)$ is therefore compatible with cancellation, not only blocking. Comparing the two isolated settings separates these explanations without assuming idle-gate invariance. Both programs are valid on all 96 legal input tuples and agree at every setting with $a = 1$, including every fully connected operational setting. In the explicit expression encodings, F uses 26 nodes and F' uses 30 across both outputs, a ratio of 1.15. This is an illustrative three-measurement record, not an additional fit to a recorded agent episode or a cohort-wide complexity-constrained search.

Recorded example. In GPT-5.6-sol's C2-R2 Neutral episode in run 1, the exact model is $\hat{F}_j = 2u + c + 1 + 2g_j s^2$. It completes source and selectivity comparisons and successful operation, but measures the isolated left output only at $s = 1$ when $u = c = 0$. An alternative retains a source effect on the isolated left output at these primary settings while leaving fully connected predictions unchanged.

$$h'_{\mathrm{L}}(u, c, s, a, b) = \begin{cases} s, & u = c = a = 0, \\ \hat{F}_{\mathrm{L}}(u, c, s, a, b), & \text{otherwise}, \end{cases} \qquad h'_{\mathrm{R}} = \hat{F}_{\mathrm{R}}. \tag{12}$$

This program fits all eight recorded measurements. At the unmeasured setting $(0, 0, 0, 0, 1)$, the model predicts $y_{\mathrm{L}} = 1$ and the alternative predicts 0. Measuring this setting would distinguish the predictions by comparing source values 0 and 1 under fixed left-path isolation (Table 30). The alternative is compatible with the record, not the actual released law. Its compiled representation uses 43 nodes.

Table 30: A blocking ambiguity in Sol's C2-R2 Neutral episode. Settings are (u, c, s, a, b) and observed readings are $(y_{\mathrm{L}}, y_{\mathrm{R}})$ in mm. The first three measurements establish source and selectivity evidence. The last is the unmeasured isolated comparison.

Setting	Observed	Model y_{L}	Alternative y_{L}
$(0, 0, 0, 1, 1)$	$(1, 1)$	1	1
$(0, 0, 1, 1, 1)$	$(3, 3)$	3	3
$(0, 0, 1, 0, 1)$	$(1, 3)$	1	1
$(0, 0, 0, 0, 1)$	Unmeasured	1	0

Size-constrained alternative search. We search the 606 source/selectivity failures and 115 blocking-only failures for alternatives using at most twice the corresponding submitted model's

syntax-tree node count. Counts sum both output expressions and exclude the policy. Each alternative is paired with the model committed before that episode's first qualifying operation and must fit the entire episode-end record while denying at least one claim under the conditions above. The search combines simplified saved witnesses with sparse arithmetic and indicator expressions over the public inputs. Numeric expressions use bounded rational coefficients, and finite-valued outputs use modular arithmetic. Blocking alternatives preserve all fully connected predictions. This offline search uses no new agent calls or simulator observations.

Table 31: Verified alternatives within twice the submitted model size. Entries are episodes with a qualifying alternative / eligible episodes (%). Size counts syntax-tree nodes in both output expressions, excluding the policy. The two failure cohorts are disjoint.

Model	Source/selectivity	Blocking only	Combined
Qwen3.8-27B	86/87 (98.9%)	—	86/87 (98.9%)
Gemini 3.8 Flash	201/203 (99.0%)	4/4 (100.0%)	205/207 (99.0%)
GPT-5.6-sol	166/167 (99.4%)	45/45 (100.0%)	211/212 (99.5%)
GPT-6 Astra	137/149 (91.9%)	66/66 (100.0%)	203/215 (94.4%)
Total	590/606 (97.4%)	115/115 (100.0%)	705/721 (97.8%)

Table 31 reports qualifying alternatives for 590/606 source/selectivity failures and all 115 blocking-only failures, totaling 705/721 episodes (97.8%). Independent verification checks every recorded measurement, the denied claim, the task's program limits on all legal inputs, and recorded interface and production outcomes. The remaining 16 episodes are unresolved by this template-based search, not proofs that no size-constrained alternative exists. These are verified existence results rather than minimum-size estimates. Parsimony may favor the submitted model, but it does not establish that the recorded experiments distinguish its causal claims from competing explanations.

F.3 A Correct Model with and without Matched Path Evidence

Table 32 compares GPT-5.6-sol and Gemini 3.8 Flash on the same C2-R2 Neutral EXPAND episode in run 1. This outcome-selected pair illustrates complete versus incomplete source and selectivity evidence. Neither trajectory contains repaired responses, invalid actions, or abnormal provider terminations. Both conversations are verified by replay against the deterministic environment.

Table 32: Actual matched trajectories. Settings are (u, c, s, a, b) and readings are $(y_{\mathrm{L}}, y_{\mathrm{R}})$ in mm.

	GPT-5.6-sol			Gemini 3.8 Flash		
Turn	Action	Settings	Reading	Action	Settings	Reading
1	`TRIAL`	$(0,0,0,1,1)$	$(1,1)$	`TRIAL`	$(0,0,0,0,0)$	$(1,1)$
2	`TRIAL`	$(0,0,1,1,1)$	$(3,3)$	`TRIAL`	$(1,0,0,0,0)$	$(3,3)$
3	`TRIAL`	$(0,0,1,0,1)$	$(1,3)$	`TRIAL`	$(0,1,0,0,0)$	$(2,2)$
4	`TRIAL`	$(0,0,1,1,0)$	$(3,1)$	`TRIAL`	$(0,0,1,1,0)$	$(3,1)$
5	`TRIAL`	$(1,0,0,1,1)$	$(3,3)$	`TRIAL`	$(0,0,2,0,1)$	$(1,9)$
6	`TRIAL`	$(0,0,2,1,1)$	$(9,9)$	`TRIAL`	$(0,0,3,1,0)$	$(19,1)$
7	`TRIAL`	$(0,0,3,1,1)$	$(19,19)$	`TRIAL`	$(1,1,1,1,1)$	$(6,6)$
8	`RUN`	$(2,1,2,1,1)$	$(14,14)$	`RUN`	$(2,0,2,1,1)$	$(13,13)$
9	`STOP`			`STOP`		

Both models submit programs equivalent to

$$y_{\mathrm{L}} = 2u + c + 1 + 2as^2, \qquad y_{\mathrm{R}} = 2u + c + 1 + 2bs^2,$$

with corrections $\pi_{\mathrm{L}} = -(1 + 2as^2)$ and $\pi_{\mathrm{R}} = -(1 + 2bs^2)$. At turn 8, each makes a correct prediction at previously untested inputs and delivers the requested target. GPT-5.6-sol predicts $(14, 14)$ and delivers $(5, 5)$ after correction. Gemini predicts $(13, 13)$ and delivers $(4, 4)$ at its different context. Both stop at turn 9 with unused action budget.

The difference is in the preceding experiments. GPT-5.6-sol changes only the source between turns 1 and 2, then isolates the left and right paths separately at turns 3 and 4. These observations supply

the two linked three-setting certificates before production. Gemini also probes both paths separately. Different source values across paths are permitted. The problem is that its fully connected trials also change primary input or context, so none can serve as the connected counterpart of its isolation trials. Turns 4 and 6 isolate a source effect on the left, but no matched gate comparison holds source, primary input, context, and the other gate fixed. Consequently, neither path certificate is complete despite the correct model and successful operation. Both trajectories lack complete four-setting evidence.

F.3.1 A COMPATIBLE ALTERNATIVE

The Gemini C2-R2 trajectory in Table 32 already has a source contrast at turns 4 and 6. Its eight measured settings nevertheless admit a different explanation of the right output. In this explanation, the right output varies with the gate settings and primary inputs, but not with the source. The saved countermodel keeps the left output equal to the submitted model and sets

$$h'_{\mathrm{R}}(u,c,s,a,b) = \begin{cases} 2u+c+1, & b=0, \\ K(u,c,a), & b=1, \end{cases} \qquad K(u,c,a) = \begin{cases} 9, & (u,c,a)=(0,0,0), \\ 6, & (u,c,a)=(1,1,1), \\ 13, & (u,c,a)=(2,0,1), \\ 0, & \text{otherwise.} \end{cases}$$

This expression agrees with the saved countermodel's right-output program on all 96 legal inputs. The countermodel fits all eight measured output pairs, including the successful production outcome. The left output still varies with s, preserving the observed source evidence. At the unmeasured $q^* = (0,0,0,0,1)$, the submitted model predicts $(1,1)$, whereas this alternative predicts $(1,9)$. This setting is also the first added `TRIAL` in the verified continuation in Table 35. Replay returns $(1,1)$, ruling out this alternative and directly linking the unresolved claim to a missing experiment. Thus evidence that an auxiliary input affects some response need not establish each claimed path. The countermodel is a witness of ambiguity in the public response language.

F.4 COMPLETION WITHIN THE REMAINING BUDGET

We test whether full *Discovery Success* can be achieved within the remaining budget for all 721 EXPAND episodes that achieve *Basic Success* but fail *Discovery Success* across the three main-evaluation runs. This cohort contains 606 episodes lacking source or selectivity evidence and 115 that complete those comparisons but lack blocking evidence. The former are the episodes in Table 36. Four models contribute eligible episodes. The other three have no EXPAND *Basic Success*. For each episode, we retain every action and observation before its final `STOP`, including invalid actions. We hold fixed the model and policy submitted for its first qualifying `RUN`. No new model or policy is synthesized. Every selected episode ends with `STOP` and still lacks complete blocking evidence at episode end.

Let Q contain the previously measured intervention tuples. Using only the public contract and the fixed submitted programs, we enumerate all valid left and right evidence certificates L and R, and all unmeasured tuples r with $a = b = 1$ where the fixed policy can complete a successful operation. Each path certificate contains the four settings required for source, selectivity, and blocking evidence (Appendix D.3). The minimum number of additional actions is

$$k^* = \min_{L,R,r} \left| (L \cup R) \setminus \big(Q \cup \{r\}\big) \right| + 1. \tag{13}$$

Every missing setting except r is measured by `TRIAL`, followed by a `RUN` at the previously untested input r. The final $+1$ counts this operation, whose outcome may complete the evidence. Measurements can support both path certificates, and each certificate includes a source contrast. `TRIAL` may isolate either path, while the final `RUN` connects both paths and meets the service and output targets. If no unmeasured, fully connected operating tuple remains, there is no valid completion without changing the recorded history before `STOP`.

Equation 13 is exact for the fixed program pair in the stationary environment. Every valid continuation must measure one pair of certificates and execute a successful operation at previously untested inputs. Measuring each missing tuple once attains that lower bound. Query order, repeated measurements, and extra non-measuring actions cannot reduce it. Enumeration covers every certificate

pair, with an independent brute-force check on small test cases. A separate verifier replays the history before STOP and each feasible continuation in the unchanged simulator, checking predictions against an independent arithmetic implementation and the full discovery criterion. Private simulator state is used for this verification, not to plan queries.

If STOP was action T under budget B, the continuation has $B-(T-1)$ actions left. We also test the conservative allowance $B-T$, without refunding STOP. The reported cost ends at the successful RUN. An additional closing STOP is executed when needed, and every feasible replay terminates legally within the original budget. Completion counts are 715 with the pre-STOP allowance and 709 with the conservative allowance.

Table 33: Full *Discovery Success* completion with fixed submitted programs. Episodes achieve *Basic Success* but fail *Discovery Success* in EXPAND, pooled across three runs. Completed counts use the pre-STOP budget. Actions give median (range) over finite completion costs, including a successful RUN at previously untested inputs but excluding a closing STOP.

Model	Episodes	Completed	Actions	No STOP refund
Qwen3.8-27B	87	87	4 (3–5)	87
Gemini 3.8 Flash	207	206	3 (2–4)	205
GPT-5.6-sol	212	212	3 (1–5)	212
GPT-6 Astra	215	210	3 (2–4)	205
Total	721	715	3 (1–5)	709

Table 33 shows that 715/721 episodes (99.2%) admit an in-budget completion, with a median of 3 additional actions. Every such continuation passes replay verification. Among the source/selectivity-incomplete episodes, 605/606 can complete the full criterion. Among those missing only blocking evidence, 110/115 can do so. Under the conservative budget, the overall count remains 709/721 (98.3%).

The six exceptions appear in Table 34. Five require one more action than remains before STOP. In Astra run 3, C1-R2 Counter Prior, all 16 fully connected operating settings have already been measured. No additional action budget can supply a previously untested operating input without changing that recorded history. These continuations do not change agent scores.

Table 34: Cases without an in-budget completion. Available actions are counted immediately before STOP. n/a means that no previously untested, fully connected operating input remains.

Model	Run	Cell	Skin	Required	Available
Gemini 3.8 Flash	1	C3-R2	Neutral	4	3
GPT-6 Astra	1	C6-R1	Counter Prior	3	2
GPT-6 Astra	2	C1-R1	Counter Prior	3	2
GPT-6 Astra	3	C3-R2	Neutral	3	2
GPT-6 Astra	3	C1-R2	Counter Prior	n/a	1
GPT-6 Astra	3	C1-R1	Counter Prior	3	2

F.4.1 COMPLETING THE RECORDED GEMINI EPISODE

Continue the same Gemini episode from Table 32, before its turn-9 STOP. Its exact model and policy were submitted at turn 8 and remain fixed. Table 35 shows the recorded minimum-cost continuation for full *Discovery Success*. Four additional actions complete source, selectivity, and blocking evidence at $u=c=0$. The left-path certificate uses turns 5, 9, 11, and 12 with $b=1$. The right-path certificate uses turns 1, 5, 9, and 10 with $a=0$. Both compare source values 0 and 2 while holding the opposite gate fixed.

The fourth added action is a RUN at previously untested inputs with both paths connected. Its predicted raw output is $(1,1)$, and corrections $(-1,-1)$ deliver the target $(0,0)$ with service receipts $(0,0)$, as required at source 0. Its measurement completes the left-path certificate and *Discovery Suc-*

Table 35: Verified offline continuation of the Gemini trajectory in Table 32, with its submitted programs fixed. Four planned actions replace its turn-9 STOP and complete *Discovery Success*.

Turn	Action	(u, c, s, a, b)	Raw outputs	Correction
9	`TRIAL`	$(0, 0, 0, 0, 1)$	$(1, 1)$	n/a
10	`TRIAL`	$(0, 0, 2, 0, 0)$	$(1, 1)$	n/a
11	`TRIAL`	$(0, 0, 2, 1, 1)$	$(9, 9)$	n/a
12	`RUN`	$(0, 0, 0, 1, 1)$	$(1, 1)$	$(-1, -1)$

cess at turn 12. A closing `STOP` uses turn 13, within the 20-action budget even without refunding the original stopping action. The offline planner uses the fixed programs and recorded measurements.

F.5 Recorded Explanations and Comparisons

We inspect the model-authored explanations alongside the measured comparisons in the same 722 episodes used in Table 28. The comparison diagnostics below concern the 606 episodes with *Basic Success* but incomplete *Path Tests*. Table 36 gives the counts and run-to-run variation for the parenthesized percentages in Figure 3.

A matched gate comparison uses two valid `TRIAL` or `RUN` measurements that differ only in one gate, with the same positive source value. For the left gate, u, c, s, b are fixed and a changes. For the right, u, c, s, a are fixed and b changes. Coverage depends only on the measured settings, not on output values or the order of the actions. Pairs are never assembled across episodes or repeated runs. A pair may be present without the agent explicitly describing it as a comparison. Table 37 gives the pooled counts underlying the three-run rates in Table 36.

Table 36: Matched gate comparisons among Expand episodes with Basic Success but incomplete Path Tests. Rates (%) are three-run mean $\pm$ sample standard deviation.

Model	Episodes (n)	No comparison	One gate	Both gates
Qwen3.8-27B	87	48.0 ± 3.1	52.0 ± 3.1	0.0 ± 0.0
Gemini 3.8 Flash	203	64.0 ± 2.9	31.5 ± 1.7	4.5 ± 1.6
GPT-5.6-sol	167	70.6 ± 13.1	7.8 ± 2.1	21.6 ± 11.1
GPT-6 Astra	149	79.1 ± 6.8	6.0 ± 0.1	14.8 ± 6.7

Table 37: Matched gate comparisons in the failure cohort. Counts pool three runs.

Model	Episodes (n)	No comparison	One gate	Both gates
Qwen3.8-27B	87	42	45	0
Gemini 3.8 Flash	203	130	64	9
GPT-5.6-sol	167	118	13	36
GPT-6 Astra	149	118	9	22
Total	606	408	131	67

Matched settings and selective effects are distinct. Of the 67 episodes covering both gates, some comparisons leave the tested output unchanged or also change the other output. We separately apply the response-dependent source and gate conditions to identify the first missing evidence requirement. We check both selective gate effects first, both connected source contrasts second, and their shared settings last (Table 38). These categories are mutually exclusive by construction. An episode in the first category may also lack source comparisons, so the table does not claim that only one requirement is missing. In the last category, both comparison types exist for both outputs, but they do not form complete path certificates.

Table 39 separates isolation trials from matched evidence. Isolation trials require each path to be measured with $s > 0$ while isolated and the other path connected. Source comparisons require each output to change across two source values with all other inputs fixed and its path connected. Gate comparisons require changing only each output's own gate to change that output but not the other.

Table 38: First missing evidence requirement, checked in column order. Counts pool the same failure cohort as Table 36.

Model	Episodes (n)	Gate effects	Source effects	Shared settings
Qwen3.8-27B	87	87	0	0
Gemini 3.8 Flash	203	199	3	1
GPT-5.6-sol	167	133	31	3
GPT-6 Astra	149	127	12	10
Total	606	546	46	14

These source and gate comparisons need not share a connected setting, so their separate presence does not imply complete *Path Tests*.

Table 39: Measured comparisons among the 606 analyzed episodes lacking source or selectivity evidence across three runs. Counts are episodes. Each comparison column requires coverage of both outputs.

Model	Failures	Isolation trials	Source comparisons	Gate comparisons
Qwen3.8-27B	87	18	60	0
Gemini 3.8 Flash	203	125	129	4
GPT-5.6-sol	167	167	34	34
GPT-6 Astra	149	149	86	22

Every analyzed Astra and Sol failure includes both active isolation trials. Their missing evidence therefore cannot be attributed simply to never trying isolation. The decomposition instead places 546 failures before complete selective gate evidence, 46 after that evidence but before complete source contrasts, and 14 at the connection of those comparisons.

We retain all 8149 per-action explanation records from the eligible episodes, including missing fields. Keyword matching for assumption language retrieves 44 candidates, including statements accepting or rejecting an assumption. The illustrative explanations below are checked against preceding recorded actions and measurements. They document stated rationales, not internal reasoning, and do not determine benchmark scores.

In Qwen 27B's run-2 C2-R3 Neutral episode, the turn-6 explanation states, "Since a=1 in all these, it's hard to tell if a matters," then adds, "Let's assume the extra offset is a * min(source, 1)." At that point, neither path has been isolated at an active source setting. A source contrast has been measured, but gate dependence remains untested. The agent subsequently measures the left-isolated setting at turn 7, then completes a qualifying `RUN` at turn 9. It never measures an active right-isolated setting and does not complete either matched path certificate. The record shows an explicit structural assumption followed by partial testing, not a complete absence of experimentation.

In Astra's run-1 C3-R3 Neutral episode, the final explanation states, "Swapped isolation trials established the auxiliary effect independently on both paths." Turns 4 and 5 do swap $(a, b) = (0, 1)$ and $(1, 0)$ at fixed $(u, c, s) = (1, 0, 1)$. Both gates change, however, so neither measurement isolates one gate's effect while holding the other fixed. The episode establishes source dependence and achieves *Basic Success*, but contains no selective single-gate comparison. Its explanation cites experimental support rather than explicitly declaring a structural premise.

F.6 Family Differences in Comparison Completion

We analyze comparison completion in all 1,512 `EXPAND` episodes in the seven-model, three-run panel. Table 40 examines Astra, which achieves *Basic Success* and performs both active isolation trials in every family. Every completed pair of source/selectivity certificates precedes a qualifying `RUN`. The C5–C6 path-evidence gap occurs despite exact models and active isolation trials in both families.

Table 40: Astra comparison completion by family (%). Each family contains 36 episodes across three runs. Gate comparisons require a selective matched contrast for each output, without requiring that it joins a source contrast.

Family	Source Test	Gate comparisons	Path Tests
C1	91.7	36.1	36.1
C2	75.0	44.4	30.6
C3	86.1	38.9	16.7
C4	80.6	25.0	22.2
C5	52.8	30.6	8.3
C6	100.0	72.2	72.2

C5 has both fewer matched source comparisons and fewer selective gate comparisons than C6. Some episodes also collect these comparisons at settings that do not form a shared connected measurement for a path certificate. For C5, selective gate comparisons are present in 30.6% of episodes, but only 8.3% complete both path certificates. For C6, all episodes establish source dependence, and both rates are 72.2%. This localizes the observed gap to comparison completion. Output arithmetic, primary-input domains, source functions, and narrative differ together across families, so these data do not isolate an arithmetic or domain effect.

Leaving out each family in turn gives Astra *Path Tests* between 22.8% and 35.6%, compared with 31.0% on the full release. Each retained mechanism keeps all three skins and all three runs. This range describes sensitivity to the composition of the fixed benchmark, not a confidence interval for unseen mechanisms. The reported run standard deviations describe repeated execution on the same release.

G Asymmetric Cross-Path Extension

This extension tests the correctness–evidence distinction in a second causal structure. The source affects both outputs, but disconnecting the left path also diverts influence to the right. The claims and evidence criteria therefore change from selective independent paths to source-dependent cross-path interaction. The main benchmark and its scores are unchanged.

The extension uses the numerical mechanical (C2) and fluid (C5) families. With $q = (u, c, s, a, b)$, the response law is

$$\begin{aligned} y_{\mathrm{L}}(q) &= m_{\mathrm{L}}(u, c) + \delta + ah(s), \\ y_{\mathrm{R}}(q) &= m_{\mathrm{R}}(u, c) + \delta + bh(s) + k(1 - a)bh(s). \end{aligned} \tag{14}$$

Here $m_{\mathrm{L}} = m_{\mathrm{R}} = 2u + c$ for C2 and u for C5. Writing $d = 1$ for C2 and $d = 1/2$ for C5, we set $\delta = d$ and $h(s) = d\gamma f(s)$. The source function is either $f(s) = s$ or $f(s) = \mathbf{1}[s = 3]$, with source gain $\gamma \in \{1, 2\}$ and rerouting gain $k \in \{1, 2\}$. The two families, two functions, and two values of each gain give 16 parameterized instances of one asymmetric structure. All use the Neutral narrative and Expand variant.

The inputs are $u \in \{0, 1, 2\}$, $c \in \{0, 1\}$, $s \in \{0, 1, 2, 3\}$, and $a, b \in \{0, 1\}$, giving 96 legal settings and 24 fully connected operational settings. At $a = b = 1$, Equation 14 agrees with the corresponding independent-path law. At $a = 0, b = 1$, the right source term increases to $(1+k)h(s)$. Thus, correct connected operation alone does not test the rerouting account, whereas exact model checking includes these isolation settings.

We evaluate Astra, Sol, and Gemini in three independent conversations per instance, for 144 episodes. The action budget is 20 and the response cap is 4,096 tokens, including reasoning, with one same-cap formatting repair. Model identifiers, reasoning effort, and decoding settings match Appendix D.4. Only Direct is evaluated. Neither the rerouting law nor its parameters or evidence certificates are given to agents. The reference below uses a declared structural prior that is not supplied to the evaluated models.

The prompt describes a shared manifold and explicitly separates command delivery from response properties. The following excerpts reproduce the apparatus descriptions and common control and evidence instructions, with display-only line wrapping. The remaining interaction and expression

protocol follows Appendix D.1. A `RUN` requires $a = b = 1$, including at $s = 0$. Full prompts are retained with the episode records.

C2 apparatus description.

```
Two indexing shafts have separate displacement meters and share a service-air
manifold. The manifold accepts a four-position pressure command. Each branch
has its own connection control.
```

C5 apparatus description.

```
Two dosing channels have separate collection meters and share a wash-fluid
manifold. The manifold accepts four bolus settings. Each branch has its own
connection control.
```

Shared control description.

```
b controls the right service connection separately. The controls do not remove
a primary unit or directly edit either meter. Service receipts record a*source
at left and b*source at right. These receipts specify command delivery, not
the raw response law. They do not guarantee that changing a connection leaves
the opposite raw output unchanged or that disconnection eliminates every
influence of the source on a raw output.
```

Shared evidence instruction.

```
Physical control alone is not a discovery certificate. Evidence for extending
the manual's causal scope must distinguish an auxiliary effect from an error
in the primary relation and experimentally support the explanation of how the
influence acts through the connection controls on both outputs.
```

The apparatus wording and the description of the required path account differ from the main Direct prompt, as do the response laws and evidence criteria. The extension tests the evaluation approach on changed causal claims, rather than estimating a topology-only effect on performance.

For evaluation, *Basic Success* requires an exact model over all 96 legal settings, a valid policy over all 24 operational settings, and a correctly predicted, successful fully connected `RUN` at a previously unmeasured input tuple. The model and policy must be committed before that outcome, and at least one earlier measurement is required. *Discovery Success* additionally requires the four comparisons below by the same qualifying `RUN`, including its measurement. The evaluator searches all measured combinations. Different claims may use different (u, c) settings and source pairs.

Fix (u, c) and abbreviate the response as $y_j(a, b, s)$.

1. **Source dependence.** Two distinct source values s_0, s_1 at $a = b = 1$ change both outputs.
2. **Source-dependent joint-gate interaction.** Measure all eight settings of $(a, b, s) \in \{0, 1\}^2 \times \{s_0, s_1\}$ and require

$$\begin{aligned} &J(s) = y_{\mathrm{R}}(0, 1, s) - y_{\mathrm{R}}(1, 1, s) - y_{\mathrm{R}}(0, 0, s) + y_{\mathrm{R}}(1, 0, s), \\ &J(s_0) \neq J(s_1). \end{aligned} \tag{15}$$

   The comparison distinguishes a source-dependent joint-gate effect from a source-independent gate interaction or lower-order additive effects.
3. **Left-path blocking.** With $b = 1$, a nonzero left source contrast at $a = 1$ vanishes at $a = 0$. Both source values are measured in both gate states.
4. **Right-path blocking under rerouting.** With $a = 0$, a nonzero right source contrast at $b = 1$ vanishes at $b = 0$, again using four measured settings. This tests blocking when the rerouted influence is present.

For Equation 14, $J(s) = kh(s)$. Any completed eight-setting interaction comparison therefore also contains the other three certificates. Their component rates describe different missing comparisons but are not independent outcomes. Each certificate concerns its measured settings, rather than verifying the response law over the entire domain.

The reference assumes the affine cross-path form in Equation 14, with unknown residual, source function, and rerouting gain. At $u = 1, c = 0$, it measures all four source values with both gates connected, identifies an active source from these observations, and measures the six missing gate/source settings at that source and at $s = 0$. These ten measurements determine δ, h, and k under the prior

and complete the evidence. A fully connected `RUN` at previously untested $u = 0, c = 1$ and a final `STOP` bring the total to 12 actions. Replay verifies *Discovery Success* on all 16 instances.

Table 41 reports per-run counts and discovery conditional on *Basic Success*. Repeats use the same 16 instances. The standard deviations in Table 1 describe run-to-run variation, not uncertainty over independently sampled structures.

Table 41: Per-run counts in the asymmetric extension. Each run has 16 instances. The last column pools the three runs and conditions on Basic Success.

Model	Basic counts	Discovery counts	Discovery / Basic
Astra	16, 16, 16	2, 1, 1	4/48 (8.3%)
Sol	15, 14, 15	1, 0, 1	2/44 (4.5%)
Gemini	13, 12, 12	0, 1, 0	1/37 (2.7%)

Table 42: Evidence completion in the asymmetric extension. Each count is out of 48 episodes and uses the full episode record.

Model	Source	Joint gates	Left block	Right block
Astra	48	4	36	7
Sol	25	2	15	8
Gemini	46	1	31	1

The episode-end counts in Table 42 retain the gap in joint-gate evidence even after all remaining agent actions. Astra records source dependence in every episode and left blocking in 36, but joint-gate interaction in only four. Sol records each of these comparisons less consistently. Gemini records source dependence in 46 episodes and left blocking in 31, but joint-gate interaction and right blocking in only one. All seven episodes with complete interaction evidence also achieve *Discovery Success*, so late evidence does not account for the remaining failures. Gemini reaches the response limit in 24 of 48 episodes, so its aggregate failure rate does not isolate evidence selection from response limitations.

To examine the unresolved claims, we analyze all 122 episodes with *Basic Success* but without *Discovery Success*. For each episode, we use the first qualifying model and all valid measurements through episode end. Candidate programs must reproduce every recorded raw output, be defined on all 96 legal settings, and satisfy the task's expression limits of 512 nodes and depth 20. The untargeted output is held equal to the submitted model. Alternative existence is a witness to unresolved claims, not the rule used to score discovery.

We construct three restricted kinds of alternative. A source-denying alternative makes one output independent of s at every fixed (u, c, a, b). A joint-gate alternative makes $J(s)$ constant across source values at every fixed (u, c), while permitting gate offsets and lower-order interactions. A blocking-denying alternative retains the submitted model outside the relevant isolated state and makes the isolated source contrast nonzero for every source pair with an observed nonzero contrast while that path is connected. The relevant gate states are $(a, b) = (0, 1)$ for left blocking and $(0, 0)$ for right blocking. Thus, blocking alternatives preserve every fully connected prediction. Without an observed contrast while that path is connected, we do not assess blocking alternatives for that path.

Table 43: Episodes admitting a claim-denying alternative in the asymmetric extension. Source counts an alternative for either output. Columns may overlap.

Model	Incomplete	Source	Joint gates	Left block	Right block
Astra	44	0	44	12	32
Sol	42	11	42	7	24
Gemini	36	0	36	7	36

All 122 episodes admit an executable alternative with source-independent J, even though the submitted models correctly encode the source-dependent interaction. Table 43 also reports alternatives

denying source dependence or blocking. These witnesses fit the complete measurement record, not only the observations preceding the first successful operation.

The search solves equality and inequality constraints on the finite response table, preferentially minimizing the number of changed entries relative to the submitted model and then their absolute changes. It compiles the resulting table using four variable orders and checks the compiled program over every input. A timeout or failure to meet expression bounds is reported as unknown, not as exclusion. No search in this panel returned unknown. The search does not minimize syntax-tree size or impose a ratio to the submitted model's size. Table 44 reports constructed sizes and search statuses. The compact-alternative result for the main benchmark (Appendix F.2) is separate from this extension.

Table 44: Alternative-search status and compiled size in the asymmetric extension. Statuses count five searches per incomplete episode, with source dependence tested separately for each output. Nodes and Size ratio are medians over constructed alternatives. Ratios use the corresponding submitted model.

Model	Found	Excluded	No anchor	Unknown	Nodes	Size ratio
Astra	88	123	9	0	231.5	9.08
Sol	92	84	34	0	207.0	6.74
Gemini	79	100	1	0	234.0	8.86

For the completion analysis, we hold the model and policy from the first qualifying `RUN` fixed, retain the episode history, and replace a final valid `STOP` with further actions. Since every discovery success requires an eight-setting interaction comparison, and each valid interaction comparison in the exact model contains the other certificates, enumerating all such comparisons is sufficient for an exact minimum-cost search. For each one, we enumerate every still-unmeasured fully connected setting for the final `RUN`. Its measurement can also fill a missing entry of the comparison. The cost is the number of missing `TRIAL`s plus this final `RUN`, with no additional `STOP` required for success.

Table 45: Fixed-program completion in the asymmetric extension. Incomplete episodes achieve Basic Success but not Discovery Success. Actions gives the minimum additional cost as median [range], including plans exceeding the remaining budget. All in-budget plans were replayed successfully.

Model	Incomplete	Alternative	In budget	Over budget	Actions
Astra	44	44	35 (79.5%)	9	3 [1–4]
Sol	42	42	41 (97.6%)	1	3 [1–5]
Gemini	36	36	29 (80.6%)	7	3 [2–5]

All 122 episodes admit a continuation using unchanged programs. The exact minimum is 1–5 additional actions, with median 3. In 105 episodes (86.1%), this fits the remaining action budget, and independent simulator replay confirms completion of all evidence and successful operation. The other 17 episodes require more actions than remain, comprising nine Astra, one Sol, and seven Gemini episodes. They remain in the denominator. Thus, the extension preserves the distinction between correct programs and completed experimental support, while quantifying when the missing evidence could still have been collected within the original budget.

H Limitations and Discussion

1. **Fixed structure and scope.** The main benchmark deliberately holds one two-path structure fixed across six families, keeping interventions and claim-specific evidence criteria comparable across response laws. The asymmetric extension examines the correctness–evidence distinction under source-dependent cross-path interaction. These two structures support a focused diagnosis of evidence completion. Broader topologies require intervention access and evidence criteria appropriate to their causal claims.
2. **Controlled experimental conditions.** Finite controls and deterministic, stationary responses permit exact model checking and direct comparison of competing predictions.

They also let us verify whether missing experiments can be completed within the remaining action budget without changing the submitted programs. Extending this evaluation to noisy environments requires repeated measurements and uncertainty estimates, while feedback and delayed effects require comparisons over trajectories.

3. **Correctness and experimental support.** *Discovery Success* requires exact modeling and successful operation together with controlled evidence for specified causal claims at tested settings. Model correctness is checked over all legal inputs, while the evidence criterion assesses completion of claim-specific experiments rather than exhaustive experimental coverage of the input domain. The distinction separates what the evaluator can establish about a model from what the agent has experimentally supported through its own measurements.