EPISTEMIC CONTROL FOR OPEN-WORLD AI SCIENTISTS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-scientist systems increasingly automate hypothesis generation, experimentation, and analysis. Yet predictive success alone does not determine whether a scientific explanation should be retained, revised, or replaced. We study epistemic control: the decisions connecting evidence acquisition to accepted explanations and scientific claims. We organise the problem around representation revision, abductive hypothesis generation, and the control of inquiry and belief revision, with six operational criteria for their coordination. We investigate selected criteria using a literature-conditioned discovery prototype and separate controlled policy studies. The prototype acquires variables from a predefined catalogue and generates expressions within a fixed executable language; the policy studies control candidates, evidence access, and revision opportunities. The results show that evidence-selection and revision policies affect inquiry efficiency and accepted conclusions, that a genuine cause can look redundant for prediction, and that recovering the right variables need not recover the generating mechanism. These findings identify limitations of predictive criteria for scientific decisions and provide controlled tests of components of epistemic control. Autonomous diagnosis and revision of an inadequate scientific representation remains the broader challenge.

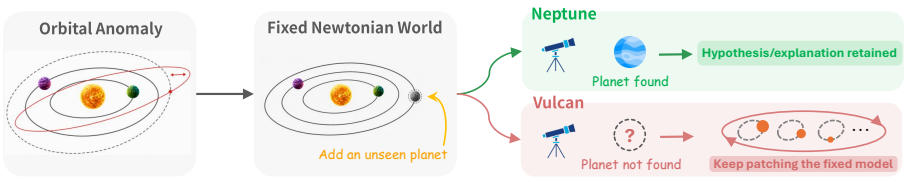
1 INTRODUCTION

In 1846, Urbain Le Verrier explained discrepancies in the orbit of Uranus by positing an unobserved planet; Neptune was found close to the predicted location. A similar proposal for Mercury's anomalous perihelion, the planet Vulcan, did not succeed; the anomaly was eventually explained by general relativity (Le Verrier, 1846; Roseveare, 1982; Einstein, 1915). These cases illustrate a recurring scientific decision: when observations challenge an explanation, should a scientist introduce another hypothesis within the current theory, reconsider the theory's assumptions, or gather more evidence before deciding? An improved fit alone does not settle that question.

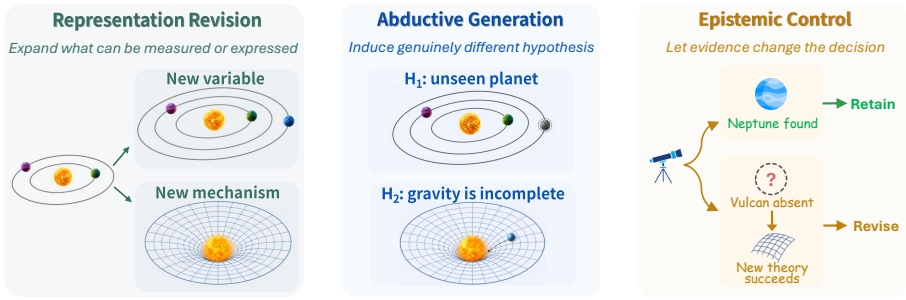
AI-scientist systems increasingly automate literature search, hypothesis generation, experimentation, analysis, and writing (Lu et al., 2026; Gottweis et al., 2026; Ghareeb et al., 2026; Aygün et al., 2026). Their scientific conclusions depend on decisions connecting these activities: which uncertainty to investigate, what the resulting evidence supports, and whether an accepted explanation should be retained, revised, or withdrawn. We call the coordination of these decisions **epistemic control**. Our focus is the choice of what to revise, especially when an agent must consider whether its available variables or model class are themselves inadequate. Prediction error is one reason to reconsider an explanation; inconsistent assumptions and uncertainty about where a model applies can also motivate inquiry.

This choice requires distinguishing the hypotheses an agent has considered from those its scientific representation can express. We describe that representation as $\mathcal{R} = (\mathcal{V}, \mathcal{L})$, comprising a scientific vocabulary $\mathcal{V}$ and an executable hypothesis language $\mathcal{L}$ (Section 2.2). Failure to find an adequate hypothesis may reflect insufficient search, an inadequate representation, or evidence that does not yet distinguish the alternatives. These possibilities motivate three complementary capabilities: **representation revision** changes the available vocabulary or hypothesis language; **abductive hypothesis generation** proposes alternative explanations; and **epistemic control** governs inquiry and the conclusions drawn from it. When a prediction fails, the error alone does not tell the agent which of these responses is called for. We use **open-world scientific discovery** for the broader setting in which

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.



(a) A failure case in closed-world scientific discovery, where local fit can preserve the wrong scientific world



(b) The capabilities open-world scientific discovery should require

Figure 1: Similar anomalies can require different scientific responses. (a) The Neptune and Vulcan cases motivate the choice between retaining an auxiliary hypothesis and reconsidering the theory; continued patching illustrates a possible failure mode. (b) The three complementary capabilities that organise our framework; we do not claim that this decomposition is the only possible one. The historical examples motivate the broader problem; our experiments test selected evidence-selection and revision decisions, and autonomous theory change is outside their scope.

the scientific representation itself remains open to revision[1]; the question, evidence sources, and acquisition budget may still be bounded.

Mechanistic World Models emphasise explanatory representations and active inquiry, and identify mechanism management (when to add, refine, or reorganise mechanisms, and what evidence to seek) as an open challenge (Posner et al., 2026). Work on abduction emphasises the generation of new explanatory premises through grounded reasoning (Zahavy, 2026). We address selected decisions within this broader agenda through explicit implementations and controlled experiments. Drawing on hypothesis generation, experimental design, hypothesis testing, and belief revision (Section 5), we make three contributions:

i) **An operational framework for epistemic control.** We formulate six criteria connecting proposed explanations to evidence acquisition, evidential support, and the retention, revision, or withdrawal of accepted claims (Section 2.3).
ii) **Implementations and controlled evaluation tasks.** We implement selected criteria in a *discovery prototype*, an agent that reads a scoped literature corpus, chooses which quantities to measure, and fits executable hypotheses to the resulting data, and in separate *policy studies* that fix the candidate explanations and the available evidence, so as to test evidence selection, failure-specific responses, and revision of accepted explanations (Section 3).
iii) **Experiments showing why accurate prediction is not enough.** Two agents given the same evidence can both make correct predictions, yet only one withdraws an explanation contradicted by that evidence. Other experiments show that predictive deletion rules can discard genuine causes, and that predicting a target need not recover the intermediate mechanisms that produce it (Section 4).

2 Epistemic Control in Scientific Discovery

2.1 What Can Require Revision?

A failed prediction challenges the hypotheses and auxiliary assumptions used to derive it, but does not by itself reveal which of them is at fault (Duhem, 1954; Quine, 1951). The candidate may be wrong, the search may be insufficient, a relevant quantity may be unavailable, or the model class may exclude an adequate explanation. Measurement problems, noise, and observational equivalence (two

[1]Distinct from the open-world assumption in knowledge representation, where the vocabulary is fixed and facts absent from the knowledge base may nevertheless be true.

Table 1: Operational criteria for epistemic control and its supporting capabilities. Glyphs indicate representation revision (⧫), hypothesis generation (•), and epistemic control (★).

Criterion	Decision	Example failure
P1 Representation revision (⧫)	What changes, and how could its consequences be tested?	Renaming a quantity without new explanatory or measurable content
P2 Failure diagnosis (★)	Which explanation of the failure should be investigated?	Applying the same repair to every discrepancy
P3 Hypothesis generation (•)	Which distinct explanation or testable implication should enter consideration?	Searching an incomplete list of candidate hypotheses indefinitely
P4 Evidence acquisition (★)	Which action addresses relevant uncertainty under the budget?	Repeating supportive observations while a discriminating test is available
P5 Evidential support (★)	What does the evidence justify for this claim and scope?	Treating predictive usefulness as sufficient evidence of causality
P6 Belief revision (★)	Which accepted claims should be retained, revised, or withdrawn?	Keeping a claim after accepting evidence against it

candidates that make the same predictions on every available observation) can also prevent a reliable conclusion. These possibilities may coexist: the appropriate response can combine model search, new measurements, and checks of existing assumptions. A diagnosis is therefore itself a hypothesis to investigate; prediction error alone does not deliver it.

We distinguish three complementary capabilities. **Representation revision** (C1; ⧫) changes the variables, concepts, or model language available to inquiry. **Abductive hypothesis generation** (C2; •) proposes alternative explanations. **Epistemic control** (C3; ★) selects evidence, evaluates its implications, and governs acceptance and revision. The distinction is operational: generating an explanation does not establish its adequacy, and interpreting counterevidence does not ensure that an accepted claim is withdrawn. We do not claim that every scientific agent needs three separate modules, or that this decomposition is the only one possible.

2.2 Representations and Candidate Hypotheses

The current scientific representation is

$$\mathcal{R}_t = (\mathcal{V}_t, \mathcal{L}_t), \qquad \mathrm{Hyp}(\mathcal{R}_t) = \mathrm{Hyp}(\mathcal{V}_t, \mathcal{L}_t), \tag{1}$$

where $\mathcal{V}_t$ is a vocabulary of scientific entities, quantities, and concepts, and $\mathcal{L}_t$ specifies how they form executable models. Examples include equations, symbolic programs, structural causal models (Pearl, 2009), and dynamical systems. The hypotheses actually considered form a subset of those expressible in the language (Appendix A.1). Adding a candidate or expanding a search library within the same language changes search; changing the vocabulary or language changes the representation. Adequacy is relative to stated scientific criteria, a domain of application, and an error tolerance.

We call a representation *revisable* when the inquiry process can propose and test changes to $\mathcal{V}_t$ or $\mathcal{L}_t$. Finite search and finite evidence generally do not certify that no adequate hypothesis is expressible. The practical objective is a provisional, evidence-informed choice between continued search, representation revision, and further inquiry under the available resources. The question, the available literature, the measurements that can be requested, and the budget may all stay fixed. Appendix A.2 defines the full inquiry state.

Bayesian model criticism and experimental design provide established tools for deciding whether a model is adequate and which experiment to run next (Gelman & Shalizi, 2013; Lindley, 1956; Chaloner & Verdinelli, 1995). Posterior updating over a fixed hypothesis class does not itself construct alternatives outside that class; a failed model check can motivate a change without supplying its replacement. Our experiments examine bounded instances of evidence selection and revision, rather than resolving the general problem of model-class change.

2.3 Operational Criteria

Table 1 makes six decisions explicit, labelled P1–P6. The criteria separate proposing a change, selecting evidence, interpreting it, and updating accepted claims; they prescribe no reasoning algorithm.

A proposed explanation should specify testable implications or the additional measurement needed to test them (P1, P3). An *evidence action* is any step that adds evidence at a cost: a measurement,

a replication, sampling in a new operating range, or a query to the literature. Such an action can discriminate current hypotheses, check measurements, replicate an uncertain result, or explore limitations shared by all current candidates (P4); the evidence-selection experiments of Section 4.2 implement the discriminative case. By *evidential support* we mean how strongly the acquired evidence bears on a specific claim, given how that evidence was obtained, the assumptions made, and the conditions it covers (P5). Observational evidence can support a predictive claim without supporting a causal one, so P5 keeps the acquisition method attached to each piece of evidence and distinguishes prediction, mechanism recovery, and causal identification. P6 requires accepted claims to respond to that assessment: a claim is *retained*, *revised*, or *withdrawn* as its support changes, and an explanation whose claim has been withdrawn is no longer treated as accepted (Popper, 1959; Lakatos, 1970). Declining a proposed model update is a different decision from withholding a scientific conclusion, and the experiments record the two separately.

Appendix A.3 gives examples of what can go wrong when an agent omits one of these operations.

3 Discovery prototype and policy studies

We use two experimental components, both scored offline against synthetic ground truth. The discovery prototype acquires measurements and searches for executable hypotheses. The policy studies run in controlled test environments that supply fixed candidate explanations and fixed evidence, and isolate decisions about inquiry, revision, and claim withdrawal. Both follow the framework of Section 2; the policy-study results do not show these behaviours operating together inside the prototype.

3.1 Variables and executable hypotheses

A task supplies a question q, target Y, initially measured variables $\mathcal{O}_0$, a value table, a scoped literature corpus, and acquisition budgets. The prototype uses $\mathcal{R}_t^{\text{proto}} = (\mathcal{V}_t, \mathcal{L}_{\text{DSL}})$. Its active vocabulary $\mathcal{V}_t = \{Y\} \cup \mathcal{O}_t$ grows through measurement of quantities from a predefined catalogue; the executable domain-specific language (DSL) stays fixed within a run.

A large language model (LLM) proposes expressions; a symbolic refitter fits coefficients on training rows and selects a sparse combination of basis terms on validation rows. The basis combines terms from the proposal, terms from retained hypotheses, and a predefined library of expansions (Appendix B.5). A proposed term outside that library can still be executed because it already belongs to the DSL, so proposing it expands the search inside a fixed language and leaves the language unchanged. Default runs fit a single equation for the target. The ordered-program condition additionally fits local equations for measured intermediates and executes them to predict the target; the agent does not itself decide to change its DSL.

3.2 The prototype's discovery loop

Each outer round supplies a literature batch conditioned on the question, current model, and candidate hypotheses. Within a batch, the agent proposes one unmeasured quantity (or a pair), resolves its name against the visible variable names and their aliases, requests its values, refits, and scores the result. A measurement request returns values only, with no indication of whether the quantity is relevant; validation error decides whether the resulting model is useful. The best proposal in a batch is accepted only if it improves validation error by a declared margin; otherwise the current model is kept and the batch is recorded as producing no update. Some controlled search studies disable that improvement rule and use ranking and stopping rules instead (Appendix B.6). Algorithm 1 specifies the standard procedure.

3.3 Policy studies and information boundaries

The policy studies give both policies the same candidate explanations, the same evidence to draw on, and the same starting state; each comparison changes one thing, either the decision policy or the responses the policy is allowed to make. Some replay a fixed evidence sequence, while others let a policy choose from a fixed menu of evidence actions. These designs isolate response or selection behaviour without requiring the agent to generate the underlying scientific alternatives.

Two boundaries constrain interpretation. First, training data fit coefficients, validation data govern selection and revision, and held-out test data are evaluated once after the final state is frozen; ground truth is used only for offline scoring. Second, reopening and withdrawal are implemented in the policy studies, and the check that a reported claim stays within what its evidence can support is a

Table 2: What each experimental component implements.

Component	Implemented behaviour	Boundary
Discovery prototype	Literature-guided measurement, expression generation, refitting, model updates	Predefined quantity catalogue and fixed DSL
Policy studies	Evidence selection, failure-specific responses, withdrawal of contradicted explanations, reopening of reported claims	Supplied candidates and controlled evidence opportunities
Offline scoring of both	Parent and graph recovery, prediction, recorded decision outcomes	Ground-truth agreement does not supply online causal identification

benchmark rule and no general identification procedure (Appendix A.4). Table 2 summarises the separation; Appendix A.3 gives the full mapping.

4 Experimental evaluation

The experiments ask four questions: (i) does the policy for selecting evidence affect how efficiently decisive evidence is reached; (ii) does the policy for interpreting and acting on evidence affect accepted explanations; (iii) when can predictive criteria give the wrong structural decision; and (iv) how much do correct variables and broader hypothesis search contribute to mechanism recovery?

4.1 Experimental setup

We use controlled experiments to isolate specific decisions in the framework. For the policy studies, we replay the same scientific situation under two decision policies and compare the resulting trajectories. A *run* is one execution of the system for a particular task, random seed, and experimental condition. A *paired comparison* consists of two runs with the same task and random seed, with all non-treatment inputs held fixed, so the two runs differ only in the experimental factor under study. Each policy study compares a *prediction-guided* policy, which selects evidence and updates its model to improve predictive validation within the declared scope, with an *epistemic-control* policy that implements the criterion under test; the two differ only in the decision policy, and in the evidence-selection experiments only in one instruction paragraph (Appendix F.1). The appendices keep the vocabulary of the released experimental record; a glossary at the start of the appendices maps it to the terms used here.

Unless stated otherwise, runs use DeepSeek-R1-Distill-Qwen-32B with three prespecified seeds per task. Budgets are matched within each comparison except when the resource itself is the treatment.

We use four evaluation sets. A ten-scenario reference set (60 runs) measures variable and mechanism recovery. Eight purpose-built tasks, with a reviewed corpus of 144 synthetic documents, each require one mechanism family: interactions, thresholds, piecewise and nested-ratio responses, alternative functional forms, or multi-step structure. Five paired policy studies (108 runs, 54 paired comparisons) isolate evidence selection and belief revision. Finally, an auxiliary-variable stress test (30 runs, 12 paired comparisons) asks whether predictive usefulness is sufficient evidence for mechanistic relevance.

We report mean squared error on a private test split that is released only after the final discovery state is frozen (held-out test MSE), together with paired outcomes and variable or graph recovery where synthetic ground truth is available. Predictive accuracy and structural recovery are reported separately. All evaluation criteria were fixed before the final runs, and failed criteria are reported unchanged.

4.2 Evidence selection and belief revision

All comparisons pass their prespecified experimental checks; four also include a held-out test criterion, while the diagnosis experiment, whose evidence is scripted and carries no numeric data, has no predictive test by design (Table 14).

Counterevidence changes accepted claims only under the revision policy. When both policies receive the same evidence sequence, both finish with a predictively correct model. However, only the epistemic-control policy withdraws the explanation contradicted by that evidence and completes the revision the evidence calls for, in 12 of 12 paired comparisons versus none. When decisive counterevidence arrives after a claim has already been reported as a conclusion, only the epistemic-control

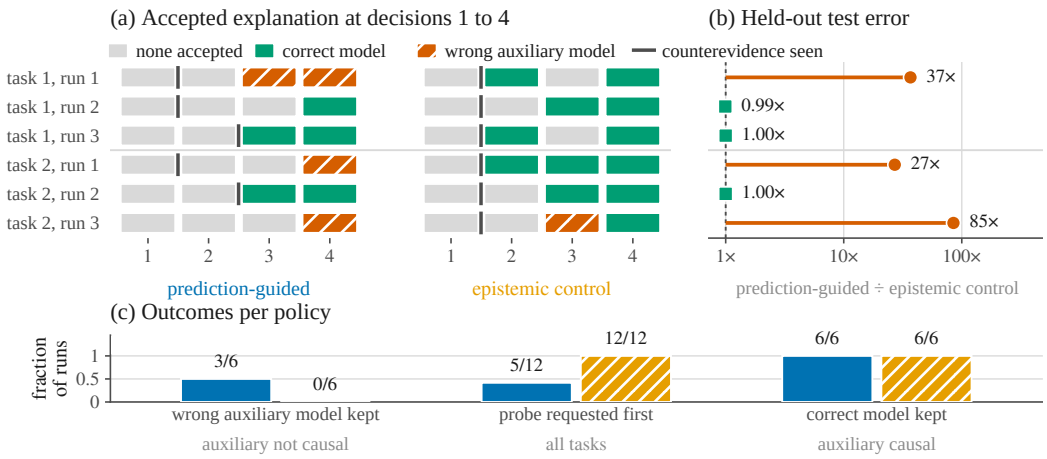


Figure 2: Final retention under different inquiry objectives. On the two tasks where the auxiliary variable is predictive but not part of the generating mechanism (tasks 1 and 2 in the figure), the prediction-guided policy keeps the wrong explanation in 3 of 6 paired comparisons, whereas the epistemic-control policy keeps it in none.

policy reopens the inquiry, withdraws the claim, and reverts the model state that depended on it, in all six paired comparisons. Thus correct final prediction does not by itself imply correct belief revision.

Discriminating evidence improves inquiry efficiency. When both policies receive the same menu of evidence actions, with identical costs, outcomes, and downstream decision rules, the policy that ranks actions by how well they separate the competing hypotheses chooses the discriminating action first in 12 of 12 paired comparisons, against 9 of 12 for the prediction-guided policy. The mean number of actions required to reach decisive evidence falls from 1.25 to 1.00. Both policies nevertheless finish correct and have identical held-out test error within every pair. With the main model this establishes an efficiency effect and no accuracy advantage. A small rerun with GPT-5.6 agrees on the efficiency direction, and there the prediction-guided policy also reaches the correct model less often, 9 of 12 against 12 of 12 (Appendix D.3); the rerun does not test the harder revision results.

Correct diagnosis does not determine the correct response. In the scripted diagnosis tasks, both policies identify the induced limitation correctly. The prediction-guided policy nevertheless attempts model revision in all six cases where the appropriate response is replication or withholding judgment. The epistemic-control policy, which conditions its response on the diagnosed failure, selects the appropriate response in every case, raising response accuracy from 0.5 to 1.0. In a separate component-revision experiment, the rates of correct revision proposals and of correct final models likewise rise from 0.5 to 1.0, because the prediction-guided policy can only add components while two of the four tasks require a removal. These experiments test response selection and revision of supplied components; they do not demonstrate autonomous discovery of what should be challenged.

4.3 When predictive fit gives the wrong structural decision

Predictive redundancy can conceal genuine causes. Across 247 leave-one-variable-out tests (the model refitted with one retained variable removed), removing the variable leaves validation error unchanged in 25 cases, so a predictive rule would mark them removable; 15 of the 25 are ground-truth parents, and 14 of those 15 come from three scenarios. In a separate paired deletion experiment, a prediction-only deletion rule removes 23 ground-truth variable instances, whereas a policy requiring explicit counterevidence removes none (Figure 8). Because this comparison generates no counterevidence, no variable ever becomes eligible for removal under the counterevidence rule. The result therefore measures the cost of prediction-only deletion against conservative retention; it says nothing about the usefulness of autonomous pruning, which the component-revision study of Section 4.2 tests separately (Appendix E.5).

The best predictor need not be part of the generating mechanism. The auxiliary-variable stress test contains two task families with similar initial predictive evidence. In both, an auxiliary measured

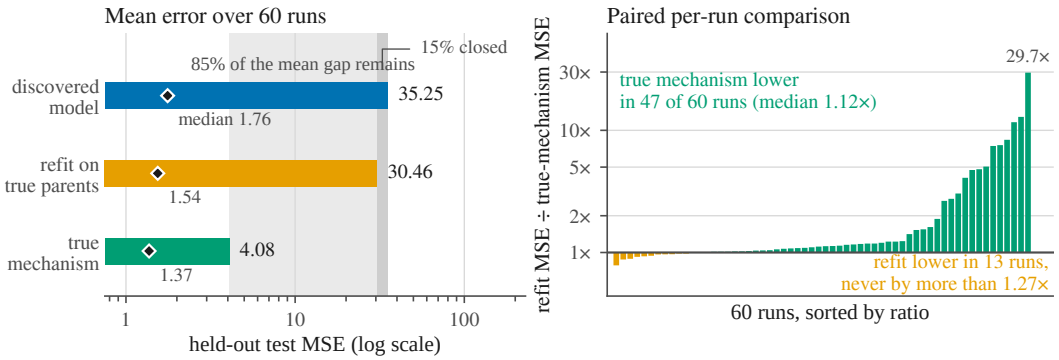


Figure 3: Variable recovery is not mechanism recovery. Left: mean and median held-out test MSE of the discovered model, the same estimator refitted on ground-truth parents, and the generating mechanism; the percentage labels give the share of the gap in mean error between the discovered model and the generating mechanism that the parent refit closes (15%) and leaves open (85%). Right: per-run error ratio of the parent-variable refit to the generating mechanism.

variable is the strongest predictor available within the initial operating range. In one family it is a genuine input to the generating mechanism; in the other it is only correlated with the target. Both policies receive the same three candidate explanations, the same five possible evidence actions, and the same total evidence budget. A decisive probe samples a wider operating range in which the correlation is broken and the mechanism is unchanged (Appendix C.4). The experiment therefore tests whether a policy distinguishes predictive association from mechanistic relevance under this one controlled shift, and nothing more general.

The two policies pursue different scientific objectives: the prediction-guided policy optimises predictive validation within the initial range, whereas the epistemic-control policy asks whether the auxiliary variable is part of the mechanism. In the six paired comparisons on the two tasks where the auxiliary variable is predictive but not causal, the prediction-guided policy keeps the wrong explanation in 3 of 6, while the epistemic-control policy keeps it in none (Figure 2). The observed difference of 0.500 meets the prespecified threshold exactly; the three discordant pairs give a one-sided sign-test $p = 0.125$. This demonstrates a policy effect and a concrete failure mode on these tasks, but provides limited evidence about how frequently such failures occur.

The epistemic-control policy requests the discriminating probe before accepting any explanation in 12 of 12 runs, compared with 5 of 12 for the prediction-guided policy, and it keeps the auxiliary variable in all six runs on the two tasks where that variable is genuinely mechanistic. When the probe outcome is supplied before the first decision, the prediction-guided policy still keeps the wrong explanation in 4 of 6 runs. Because the two policies pursue different objectives, this contrast shows what the objective does; it does not show irrational updating under a shared objective.

The prespecified predictive-dominance criterion was not met. The held-out test draws half of its rows from the initial operating range and half from the wider range. In the three paired comparisons where the policies finish with different models, the epistemic-control model has 27–85 times lower error; the remaining three are approximately tied. The prespecified requirement of at least 5% improvement in five of six pairs is therefore not met, with a result of 3/6. Because this criterion can only succeed when the baseline selects the wrong model sufficiently often, it is coupled to the baseline failure rate. We retain both the original threshold and the failed result. The final experiment followed two development versions with revised prompts and fresh seeds (Appendix E.6).

4.4 Variables, mechanisms, and hypothesis search

Recovering the right variables does not recover the mechanism. Across 60 reference runs, median held-out test MSE is 1.76 for the discovered model, 1.54 when the same estimator is refitted using the ground-truth parent variables, and 1.37 for the executed generating mechanism. Refitting on the true parents therefore closes 56% of the median gap between the discovered model and the generating mechanism; on cross-task means it closes only about 15%, because a few large-scale tasks dominate the means. These are descriptive summaries rather than a decomposition of prediction

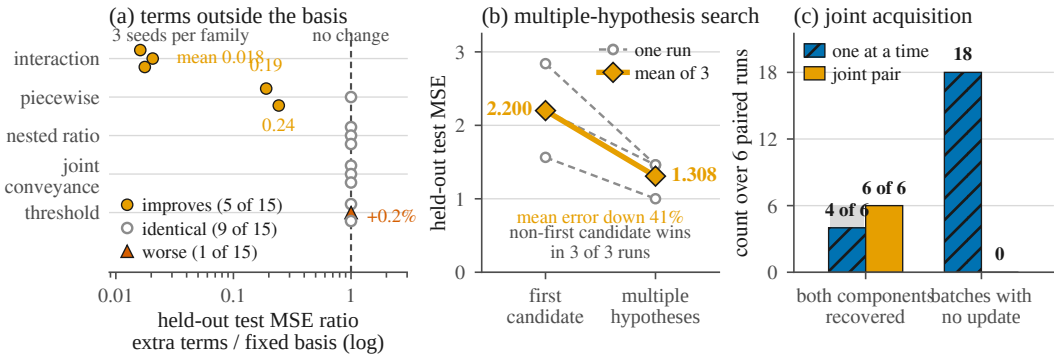


Figure 4: Controlled changes to hypothesis search. (a) Adding literature-proposed terms improves five paired comparisons; only the materials task selects a term outside the predefined search basis. (b) Multiple-hypothesis search selects a non-first candidate in all three paired comparisons. (c) Joint acquisition recovers both components in all six runs.

error into independent causes. The generating mechanism has lower error than the parent-variable refit in 47 of 60 runs (Figure 3). The generating mechanism differs from that refit in functional form and parameter values as well as in variables, so the remaining gap mixes misspecification with finite-sample estimation and regularisation effects.

Accurate target prediction does not recover the full mechanism graph. On two tasks whose mechanism runs through measured intermediate variables, all six runs that fit an ordered program (one local equation per intermediate, then one for the target) recover the benchmark graph exactly: every directed edge is present and none is extra (directed-edge F1 of 1.0, structural Hamming distance of 0). A single flat equation for the target reaches directed-edge F1 between 0.33 and 0.60, because it omits the equations for the intermediates and replaces them with direct shortcuts into the target. The aggregate held-out error ratio is 0.943, within the prespecified tolerance of 1.10; five paired comparisons improve and one is slightly worse. These results concern recovery of the supplied benchmark structure under measured intermediates and controlled representational restrictions; unconstrained causal identification is outside their scope (Figure 7, Appendix E.4).

Broader candidate search helps when the restricted search set misses a useful term. Adding literature-proposed terms improves held-out test MSE in 5 of 15 paired comparisons, ties in 9, and worsens it by 0.2% once. Only the materials task selects a term outside the predefined search basis: a product interaction that is already expressible by the fixed DSL, with mean paired error ratio 0.018 over three seeds. The crop task improves in two pairs without selecting an outside-basis term; the remaining tasks tie or incur the small loss (Figure 4a). Thus improvement from broader proposals does not by itself imply an increase in representational expressivity.

Considering more than one hypothesis can change which explanation is selected. When the first candidate is held identical and deliberately made incorrect, multiple-hypothesis search selects a later candidate in all three battery-task paired comparisons and lowers mean held-out test MSE from 2.200 to 1.308, while both conditions recover exactly the same variables. In a separate acquisition experiment, acquiring two named components jointly recovers both in 6 of 6 runs versus 4 of 6 when they are acquired one at a time, and no literature batch ends without an accepted update, against 18 such batches for one-at-a-time acquisition (Figure 4b,c). These results demonstrate benefits of broader candidate search and joint acquisition within a fixed language and predefined catalogue. They do not establish autonomous diagnosis of a language limitation or unrestricted hypothesis invention.

4.5 Robustness and Scope

Compared with a single fixed data panel, streaming acquisition (a fresh panel before each literature batch) produces 15 predictive wins in each direction across 30 paired comparisons, while recovery of the initially unmeasured parents (hidden-variable F1) falls from 0.969 to 0.893 and the share of requested measurements that are true parents (query efficiency) falls from 0.950 to 0.713. The

balanced win count does not establish predictive equivalence. The structural metrics show that incremental evidence can preserve prediction while degrading recovery of the underlying structure.

Alternating relevant and irrelevant documents across literature batches lowers hidden-variable F1 from 0.983 to 0.822 across 108 runs. This comparison does not isolate whether the effect arises from relevance, continuity, or the effective document budget. Making proxy variables (noisy readouts of hidden parents) visible from the start increases their mean share among selected non-target variables from 0.031 to 0.184 (Figure 6, Appendices E and H).

The evidence remains bounded by designed synthetic tasks, three seeds per task, and supplied scientific resources: the prototype does not invent new hypothesis languages, and the policy studies do not combine all tested behaviours in one autonomous discovery loop. The comparisons isolate particular decision policies and representational restrictions; they establish no superiority over strong Bayesian experimental-design or model-revision methods, and fresh seeds after development do not show generalisation to unseen task families. Unrestricted concept invention, autonomous choice of representation revision, and real-world mechanism discovery remain untested.

5 Related work and background

Explanatory representations and hypothesis generation. Mechanistic World Models organise knowledge around variables, reusable mechanisms, and their composition, and explicitly include hypothesis generation and experimental design (Posner et al., 2026). Causal representation learning studies variables and structures that support causal reasoning (Schölkopf et al., 2021), while symbolic regression provides executable model classes and search procedures (Brunton et al., 2016; Udrescu & Tegmark, 2020; Cranmer et al., 2020). Recent work on abduction emphasises the origin of new explanatory premises through grounded reasoning (Zahavy, 2026); our expression-search experiments do not test such concept invention. Automated discovery has a long history, including BACON, EURISKO, and compositional kernel search (Langley et al., 1987; Lenat, 1983; Lenat & Brown, 1984; Duvenaud et al., 2013; Lloyd et al., 2014). Our focus is on explicit decisions connecting such representational and generative resources to evidence and accepted claims.

Inquiry, criticism, and revision. Bayesian experimental design supplies formal objectives for evidence acquisition (Lindley, 1956; Chaloner & Verdinelli, 1995), including amortised and policy-based approaches (Foster et al., 2021; Ivanova et al., 2021; Blau et al., 2022; Rainforth et al., 2024) and methods for causal structure learning (Tigas et al., 2022; Annadani et al., 2023; Toth et al., 2022; Choo et al., 2023). The Model Discovery Agent combines LLM-assisted model discovery with Bayesian experimental design (Murphy, 2026). Our selector is a prompt-level policy and no substitute for an information-gain estimator; the current experiments do not compare it with these stronger design methods. Bayesian model criticism and debates over changing hypothesis spaces already address important aspects of revision (Gelman & Shalizi, 2013; Wenmackers & Romeijn, 2016; Howson & Urbach, 2006). We also draw on accounts of theory revision, severe testing, and unconceived alternatives (Lakatos, 1970; Kuhn, 1962; Mayo, 2018; Stanford, 2006); the criteria combine these established ideas into explicit operations and controlled tests.

Scientific agents. Robot Scientists, self-driving laboratories, and LLM systems automate substantial parts of scientific inquiry (King et al., 2009; Häse et al., 2019; MacLeod et al., 2020; Burger et al., 2020; Boiko et al., 2023; Ifargan et al., 2024; Schmidgall et al., 2025; Swanson et al., 2025; Lu et al., 2024; 2026; Ghareeb et al., 2026; Aygün et al., 2026). We do not claim that these systems lack criticism or revision; our contribution is a framework for distinguishing what an agent can express, what it considers, and what it accepts, together with controlled evidence about selected policies governing those transitions.

6 Conclusion

Scientific agents need policies for deciding what to investigate and how evidence should change their explanations. Our controlled studies show that these policies affect inquiry efficiency and revision behaviour, and that predictive criteria can preserve spurious structure or remove genuine causes in synthetic tasks. Deciding autonomously when to change the available variables or model class remains an open challenge (Appendix G).

AI USE STATEMENT

Generative AI enters this work as the object of study and as a research tool.

As the object of study. A large language model is a component of the evaluated system: it proposes quantities to measure, drafts executable hypotheses, resolves proposal names, and, in the policy studies, returns the decision objects. All confirmatory runs use DeepSeek-R1-Distill-Qwen-32B from a pinned checkpoint at temperature 0, with one transferability rerun on GPT-5.6 (Appendix D.3). Every call uses a frozen prompt template (Appendix F.1) and every generation is recorded, so these uses are measured outputs of the experiments.

As a research tool. Among tasks with required disclosure: the two synthetic literature corpora were drafted and reviewed by LLM calls and then checked by a deterministic validator (Appendix C.3); the numeric data come from author-written generators. AI coding assistants helped implement the prototype, the policy studies, and the task generators; all code was reviewed by the authors and is covered by the tests and audits of Appendix D.4. AI tools helped draft summaries of the frozen run records; every reported number is recomputed from those records by reviewed scripts, and the interpretation of every result, including the failed acceptance criteria, is the authors' own. The framework, the six criteria, the benchmark designs, and the acceptance criteria were formulated by the authors. We did not use generative AI to formulate or prove mathematical claims; translation and qualitative thematic analysis are not applicable. Among tasks with recommended disclosure, we used generative AI to edit text for readability, suggest structure, draft parts of the appendices from the frozen records, and format references; the authors wrote or revised the final text of every section. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Yashas Annadani, Panagiotis Tigas, Desi R. Ivanova, Andrew Jesson, Yarin Gal, Adam Foster, and Stefan Bauer. Differentiable multi-target causal Bayesian experimental design. In *International Conference on Machine Learning*, 2023.

Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, Marc Coram, Hao Cui, Jake Garrison, Renee Johnston, Anton Kast, Cory Y. McLean, Peter Norgaard, et al. An AI system to help scientists write expert-level empirical software. *Nature*, 654:909–916, 2026. doi: 10.1038/s41586-026-10658-6. URL https://www.nature.com/articles/s41586-026-10658-6.

Tom Blau, Edwin V. Bonilla, Iadine Chades, and Amir Dezfouli. Optimizing sequential experimental design with deep reinforcement learning. In *International Conference on Machine Learning*, 2022.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Benjamin Burger, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M. Alston, Buyi Li, Rob Clowes, Nicola Rankin, Brandon Harris, Reiner Sebastian Sprick, and Andrew I. Cooper. A mobile robotic chemist. *Nature*, 583:237–241, 2020. doi: 10.1038/s41586-020-2442-2.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. doi: 10.1214/ss/1177009939.

Davin Choo, Themistoklis Gouleakis, and Arnab Bhattacharyya. Active causal structure learning with advice. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5838–5867, 2023. URL https://proceedings.mlr.press/v202/choo23a.html.

Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Pierre Duhem. *The Aim and Structure of Physical Theory*. Princeton University Press, 1954. Translated by Philip P. Wiener.

David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning*, 2013.

Albert Einstein. Erklärung der perihelbewegung des Merkur aus der allgemeinen relativitätstheorie. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, pp. 831–839, 1915.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *International Conference on Machine Learning*, 2021.

Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013. doi: 10.1111/j.2044-8317.2011.02037.x.

Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodriques. A multi-agent system for automating scientific discovery. *Nature*, 655:497–505, 2026. doi: 10.1038/s41586-026-10652-y. URL https://www.nature.com/articles/s41586-026-10652-y.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL https://www.nature.com/articles/s41586-026-10644-y.

Florian Häse, Loïc M. Roch, and Alán Aspuru-Guzik. Next-generation experimentation with self-driving laboratories. *Trends in Chemistry*, 1(3):282–291, 2019. doi: 10.1016/j.trechm.2019.02.007.

Colin Howson and Peter Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court, 3 edition, 2006.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous LLM-driven research from data to human-verifiable research papers, 2024. URL https://arxiv.org/abs/2404.17605.

Desi R. Ivanova, Adam Foster, Steven Kleinegesse, Michael U. Gutmann, and Tom Rainforth. Implicit deep adaptive design: Policy-based experimental design without likelihoods. In *Advances in Neural Information Processing Systems*, 2021.

Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, and Amanda Clare. The automation of science. *Science*, 324(5923):85–89, 2009. doi: 10.1126/science.1165620.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

Imre Lakatos. Falsification and the methodology of scientific research programmes. In Imre Lakatos and Alan Musgrave (eds.), *Criticism and the Growth of Knowledge*, pp. 91–196. Cambridge University Press, 1970.

Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Processes*. MIT Press, 1987.

Urbain Jean Joseph Le Verrier. Recherches sur les mouvements d'Uranus. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 22:907–918, 1846.

Douglas B. Lenat. EURISKO: A program that learns new heuristics and domain concepts. *Artificial Intelligence*, 21(1–2):61–98, 1983. doi: 10.1016/S0004-3702(83)80005-8.

Douglas B. Lenat and John Seely Brown. Why AM and EURISKO appear to work. *Artificial Intelligence*, 23(3):269–294, 1984. doi: 10.1016/0004-3702(84)90016-X.

Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069.

James Robert Lloyd, David Duvenaud, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Automatic construction and natural-language description of nonparametric regression models. In *AAAI Conference on Artificial Intelligence*, 2014.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery, 2024. URL https://arxiv.org/abs/2408.06292.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651:914–919, 2026. doi: 10.1038/s41586-026-10265-5. URL https://www.nature.com/articles/s41586-026-10265-5.

Benjamin P. MacLeod, Fraser G. L. Parlane, Thomas D. Morrissey, Florian Häse, Loïc M. Roch, Kevan E. Dettelbach, Raphaell Moreira, Lars P. E. Yunker, Michael B. Rooney, Joseph R. Deeth, Veronica Lai, Gordon J. Ng, Henry Situ, Ray H. Zhang, Michael S. Elliott, Ted H. Haley, David J. Dvorak, Alán Aspuru-Guzik, Jason E. Hein, and Curtis P. Berlinguette. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6(20):eaaz8867, 2020. doi: 10.1126/sciadv.aaz8867.

Deborah G. Mayo. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, 2018.

Rowan McAllister, Yarin Gal, Alex Kendall, Mark van der Wilk, Amar Shah, Roberto Cipolla, and Adrian Weller. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 4745–4753, 2017. doi: 10.24963/ijcai.2017/661. URL https://www.ijcai.org/proceedings/2017/0661.pdf.

Kevin Murphy. Model discovery agent: LLM-assisted bayesian experiment design for data-efficient discovery of mechanistic world models, 2026. URL https://arxiv.org/abs/2608.09696.

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

Karl R. Popper. *The Logic of Scientific Discovery*. Hutchinson, 1959.

Ingmar Posner, Anson Lei, and Bernhard Schölkopf. From observation to insight: Mechanistic world models and the quest for autonomous discovery, 2026. URL https://arxiv.org/abs/2607.12474.

Willard Van Orman Quine. Two dogmas of empiricism. *The Philosophical Review*, 60(1):20–43, 1951. doi: 10.2307/2181906.

Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

N. T. Roseveare. *Mercury's Perihelion: From Le Verrier to Einstein*. Clarendon Press, Oxford, 1982. ISBN 978-0-19-858174-1.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants, 2025. URL https://arxiv.org/abs/2501.04227.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

P. Kyle Stanford. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford University Press, 2006.

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

Panagiotis Tigas, Yashas Annadani, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. Interventions, where and how? experimental design for causal models at scale. In *Advances in Neural Information Processing Systems*, 2022.

Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius von Kügelgen. Active Bayesian causal inference. In *Advances in Neural Information Processing Systems*, 2022.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631.

Sylvia Wenmackers and Jan-Willem Romeijn. New theory about old evidence: A framework for open-minded bayesianism. *Synthese*, 193(4):1225–1250, 2016. doi: 10.1007/s11229-014-0632-x.

Tom Zahavy. Position: LLMs can't jump, 2026. URL https://openreview.net/forum?id=klU4737opt. ICML 2026 Position Paper.

APPENDIX CONTENTS

A Formalising open-world representation repair **15**
A.1 Formal definitions and failure localisation 15
A.2 Discovery state and correspondence with the prototype 16
A.3 Non-interchangeability of the principles 17
A.4 Claim taxonomy and equivalence 18
A.5 Evidence and decision ledger and supporting safeguards 19

B Prototype details **20**
B.1 Literature retrieval and progressive exposure 20
B.2 Proposal, repair, and canonical resolution 21
B.3 Measurement interface and oracle semantics 22
B.4 DSL grammar, parser, and compiler 22
B.5 Hybrid symbolic refit and hypothesis management 23
B.6 No-commit, revision, and rollback 24

C Benchmarks and scientific corpora **25**
C.1 The reference panel (reference task set) 25
C.2 The capability panel (capability task set) 27
C.3 Literature construction and quality control 28
C.4 Data generation, stochasticity, and distractors 28
C.5 Matched-trajectory tasks and generators 30

D Evaluation protocol **31**
D.1 Train, validation, and private-test roles 31
D.2 Metrics and claim-level scoring 32
D.3 Matched budgets and statistical summaries 34
D.4 Leakage, correctness, and regression audits 36

E Experiment details and supplementary results **37**
E.1 Complete evidence inventory 37
E.2 Integrity and representation-failure localisation 39
E.3 Inquiry, robustness, and reference decompositions 40
E.4 Plural executable hypothesis search 43
E.5 Falsification and revision 46
E.6 Per-task results and failure traces 48

F Prompts and reproducibility **50**
F.1 Prompt and schema inventory 51
F.2 Configurations and experiment manifests 52
F.3 Models, compute, and execution cost 53
F.4 Artifact lineage and regeneration 54

G Open problems **54**

H Studies outside the main claims **56**
H.1 Proxy variables and measurement substitutes 56
H.2 The real-data feature-selection analogue 57

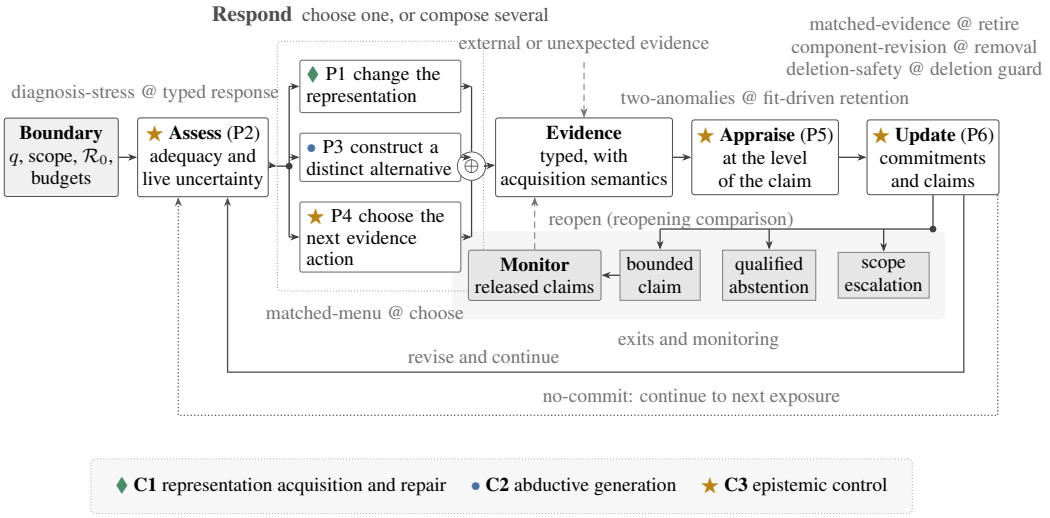


Figure 5: Conceptual scope of the epistemic-control framework. The main row is one round: the current explanation is assessed, one or more of the three responses is chosen, and the appraised evidence then updates accepted claims. The shaded band is the second level: an accepted explanation leaves the loop as a claim with stated scope (bounded claim), as an abstention, or as a request to widen the scope (scope escalation), and reported claims are monitored and can be reopened. Grey tags locate controlled studies of selected decisions; the diagram is not a single implemented loop, and representation revision and general claim monitoring remain only partially instantiated.

Vocabulary of the record. The appendices keep the vocabulary of the released artifacts, which the main text simplifies: a *cell* is one run (task, seed, arm); an *arm* is one experimental condition of a paired comparison; a *pair* is a paired comparison; a *gate* is a prespecified acceptance criterion; a *panel* is a task set or, for data, a table of row-aligned measured values; an *exposure* is one literature batch; an *account* is a candidate model or explanation; the *incumbent* is the currently accepted account; the *live set* is the set of candidate hypotheses under consideration; a *commit* accepts a trial's model update and a *no-commit* declines it while keeping the current model; the *sealed* or *private* test is the held-out test split released once after the final state is frozen; *acquisition semantics* records how a piece of evidence was obtained (observational sampling, a wider regime, a replication, or a literature passage); an *evidence action* is one costed step that adds evidence; a *harness* is the controlled test environment that runs one policy study; and *typed* repair or response means a repair or response specific to the diagnosed failure.

A Formal framework and implementation boundaries

This appendix defines the representation and inquiry state, illustrates operational distinctions among the criteria, and describes the benchmark conventions used to score claims. These definitions distinguish the general framework from the bounded implementations evaluated here.

A.1 Formal definitions and failure localisation

The definition separates the resources for expressing hypotheses from the hypotheses built with them:

$$\mathcal{R}_t = (\mathcal{V}_t, \mathcal{L}_t), \qquad \mathrm{Hyp}(\mathcal{R}_t) = \mathrm{Hyp}(\mathcal{V}_t, \mathcal{L}_t). \tag{2}$$

The vocabulary $\mathcal{V}_t$ is a typed set of scientific referents: observable or latent quantities, entities, components, categories, variable types, units, states, events, and named concepts. Membership means the item can be referred to in a hypothesis, and $\mathcal{O}_t \subseteq \mathcal{V}_t$ marks the currently measured ones; defining a concept, deciding how to measure it, and obtaining a value remain three separate events in the record.

The language $\mathcal{L}_t$ specifies how vocabulary items form executable hypotheses: its model-object families, primitives, operators, types, constraints, and execution semantics decide which scientific objects are legal, how they combine, and how a candidate produces testable consequences.

$\mathrm{Hyp}(\mathcal{R}_t)$ denotes all well-formed executable hypotheses constructible from $\mathcal{V}_t$ under $\mathcal{L}_t$. The set of candidates an algorithm actually considers is normally much smaller:

$$\mathcal{H}_t^{\mathrm{live}} \subseteq \mathrm{Hyp}_{\mathrm{search}}(\mathcal{R}_t) \subseteq \mathrm{Hyp}(\mathcal{R}_t). \tag{3}$$

The first inclusion reflects finite proposals and search; the second separates what the language can execute from what the configured generator can reach. Failure to find an expressible candidate is a search failure; exclusion of all adequate candidates by the vocabulary or language is a representational limitation. An observed failure alone does not distinguish them. This distinction prevents search expansion from being interpreted as demonstrated language revision.

Candidate search, measurement acquisition, and representation revision should be distinguished. A candidate revision $h_t \to h_{t+1}$ at fixed $\mathcal{R}_t$ changes which expressible hypothesis is under consideration. A representation revision changes the available vocabulary or language,

$$\mathcal{R}_{t+1} = (\mathcal{V}_{t+1}, \mathcal{L}_{t+1}) \neq (\mathcal{V}_t, \mathcal{L}_t), \tag{4}$$

and should be motivated by evidence for a limitation and evaluated through its consequences. Expanding a measured subset of a predefined catalogue is a bounded acquisition operation; introducing a genuinely new scientific concept is a stronger requirement. An evidence action adds evidence through the permitted interface and may test a candidate or a revision without changing the representation.

Table 3: Possible diagnoses and corresponding responses; they are not mutually exclusive or exhaustive. A poor fit alone does not identify the row.

Diagnosed limitation	**Operational criterion**	**Matched response**	**Primary object changed**
Missing vocabulary	Required entity, quantity, type, or concept is absent	Introduce or operationalise the missing item	$\mathcal{V}_t$
Inadequate language	Required model object, primitive, operator, constraint, or composition is inexpressible	Extend or replace the relevant part of the language	$\mathcal{L}_t$
Wrong expressible hypothesis	A viable correction already lies in $\mathrm{Hyp}(\mathcal{R}_t)$	Generate, edit, re-rank, reject, or replace candidates	$\mathcal{H}_t^{\mathrm{live}}$ and commitments
Evidential underdetermination	Live candidates make different predictions but current evidence does not separate them	Select a discriminative measurement, simulation, or experiment	$\mathcal{E}_{\leq t}$ through $\mathcal{I}_t$
Noise or non-identifiability	Available outcomes cannot support a stable distinction	Check measurements or withhold a conclusion	Evidence and commitment status; no automatic repair

A.2 Discovery state and correspondence with the prototype

The framework distinguishes the following state components:

$$\mathcal{S}_t = \big(\mathcal{R}_t, \mathcal{I}_t, \mathcal{H}_t^{\mathrm{live}}, \mathcal{E}_{\leq t}, \kappa_t\big). \tag{5}$$

The inquiry interface $\mathcal{I}_t$ records the actions currently permitted and their budget or availability state; $\mathcal{H}_t^{\mathrm{live}}$ is the set of candidate hypotheses under consideration. Evidence $\mathcal{E}_{\leq t}$ includes typed literature records, measurements, simulations, or interventions together with their acquisition semantics. κ_t records the accepted claims: which propositions are retained, rejected, or unresolved, the evidence for and against each, and the decision that produced each entry. It is a record of decisions and holds no hypotheses of its own.

In the prototype, $\mathcal{R}_t$ is the catalogue of variables with its measured subset plus the DSL grammar and compiler; $\mathcal{I}_t$ is the literature-batch schedule and the measurement interface with their budgets; $\mathcal{H}_t^{\mathrm{live}}$ is the set of retained candidate models; $\mathcal{E}_{\leq t}$ holds the exposed passages and the measured-value table; and κ_t is the ordered record of accepted models, declined updates, and revisions.

The standard prototype expands a measured subset of a predefined catalogue and revises hypotheses within a fixed DSL. Changes to the refit basis do not change that DSL. Broader concept invention, autonomous changes between model languages, and unrestricted SCM search remain target capabilities. The implemented commitment history is narrower than the general claim-and-dependency state described above.

Notation. Symbols follow their point of introduction: the state components of Eq. (5), the budgets of Algorithm 1, the refit objects of Appendix B.5, and the metric quantities of Appendix D.2; a complete symbol table ships with the release.

A.3 Operational distinctions among the criteria

The three-capability decomposition groups operations according to their role in inquiry. Alternative architectures may combine representation revision and hypothesis generation in one procedure or distribute epistemic control across several modules. The state representation does not determine a unique capability decomposition. Table 4 illustrates what can go wrong when a criterion is omitted while other functions remain available. These examples motivate separate evaluation questions rather than a logical independence theorem. This establishes an operational distinction, and it does not prove the set mathematically minimal or exhaustive.

Table 4: Operational distinctions among P1–P6, with their state components and relevant experimental evidence. The evidence is local to the stated comparisons; the P1 row identifies an untested stronger capability.

P	Decision governed	Failure still possible if removed	Capability; component	Measured instance
P1	What counts as a testable representation transition?	The system diagnoses a representational limit but can search only a fixed vocabulary and language.	C1; rewrites $\mathcal{R}_t$	Autonomous vocabulary or language invention is not tested; catalogue acquisition is bounded
P2	Is the account adequate, and which response fits the deficiency?	The system has repair, generation, and inquiry tools but applies an unmatched response or forces a conclusion under non-identifiability.	C3; routes diagnosis over all five components	Response accuracy 0.500 $\rightarrow$ 1.000; diagnosis is correct in both arms (§4.2)
P3	Which scientifically distinct explanation should enter consideration?	An adequate explanation is expressible but never generated; ranking and testing operate on an incomplete candidate set.	C2; rewrites $\mathcal{H}_t^{\text{live}}$	Non-first-candidate rescue 3/3, mean ratio 0.594 (§4.4)
P4	Which admissible evidence action should be selected next?	Inquiry remains generic, redundant, or confirmation-seeking and fails to resolve the consequential uncertainty.	C3; exercises $\mathcal{I}_t$	Discriminating-first action 9/12 vs 12/12 (§4.2)
P5	What does the acquired evidence establish?	Evidence is obtained correctly but interpreted with the wrong scientific criterion, such as using prediction to certify mechanism or causality.	C3; governs the reading of $E_{\leq t}$	25 of 247 exclusions marked removable, 15 ground-truth parents (§4.3)
P6	How should commitments and released claims change?	The evidence is interpreted correctly, yet obsolete commitments persist or the reported claim exceeds its evidence.	C3; rewrites κ_t	Defeated-account retirement 0/12 $\rightarrow$ 12/12 matched pairs (§4.2)

P1 describes testable representation changes; P2 concerns provisional diagnosis. P3 supplies alternatives, while P4 acquires evidence relevant to uncertainty. P5 assesses support for a claim under assumptions and scope; P6 makes that assessment effective in accepted conclusions. These functions can interact rather than occurring in a fixed sequence.

Table 5: How the six principles are instantiated in the current prototype, together with the boundary of each implementation.

	Concrete mechanism	Current boundary
P1	Canonical component activation, provisional measurement, and executable flat or ordered-program updates	Representation change is bounded by the predefined catalogue and implemented DSL; there is no general ontology or language invention.
P2	Gap-conditioned proposals, stop verification and repair, validation gain, and no-commit	Diagnosis is heuristic; the system does not yet classify failure into a general set of typed scientific causes.
P3	Multiple DSL priors, terms outside the predefined basis, hybrid symbolic refitting, joint components, and a mechanism bank	Abduction remains bounded by the executable language, available measurements, and proposal policy.
P4	Hypothesis-aware retrieval, literature-supported ranking, measurement requests, and same-exposure revisits	The prototype lacks a unified action menu and a general expected-information-gain policy.
P5	Executable validation scoring, complexity control, and local and whole-program checks	Online appraisal remains prediction-heavy; variable, mechanism, and graph criteria are primarily evaluation-layer diagnostics.
P6	Live alternatives, no-commit, synchronous state updates, guarded grouped revision, and rollback provenance	Useful autonomous deletion, general dependency propagation, and claim release remain partial; reopening and withdrawal are harness-level (Section 4.2).

A.4 Claim targets, evidence checks, and equivalence

P5 (Section 2.3) says that evidence for one kind of claim is no evidence for another; this subsection fixes the claim vocabulary, the ordering that the policy harnesses use to score reported claims, and when two hypotheses count as the same account.

Claim families and evidence. We distinguish scientific targets from the information used to select a model and from ground truth available only to the evaluator. A *predictive* claim concerns held-out performance in a stated regime; an *association* claim concerns dependence among observed quantities. A *variable* claim proposes that a quantity participates in a mechanism, a *functional* claim proposes an equation or composition, and a *graph* claim proposes an edge set. These claims can be scored against synthetic truth without being identified from the agent's evidence. In particular, survival under predictive refitting establishes selection as a predictor; it does not establish membership in the generating parent set. The benchmark target-edge comparison is

$$\widehat{E}_Y = \{(x, Y) : x \in \mathrm{vars}(h)\}, \qquad E_Y = \{(x, Y) : x \in \mathrm{pa}(Y)\}. \tag{6}$$

Its interpretation as recovery uses the evaluator's graph; the syntax of a predictive expression does not by itself identify causal direction. An *interventional* claim concerns outcomes under manipulation and requires identification under explicit assumptions. Observational evidence retains its acquisition semantics even when additional assumptions permit causal inference (Pearl, 2009). The current harness excludes interventional claims and performs no such identification analysis.

The harness claim ladder. The matched-trajectory and two-anomalies policy harnesses require every decision point to release a claim level and a one-sentence claim. They share a benchmark-specific ordering with rank function ρ. This is an implementation convention and no universal ordering of scientific knowledge: functional form and graph structure are distinct targets.

$$\begin{aligned} &\text{none} \prec \text{predictive} \prec \text{variable} \prec \text{mechanism} \prec \text{structure} \prec \text{causal}, \\ &\rho(\text{none}) = 0, \ \ldots, \ \rho(\text{causal}) = 5. \end{aligned} \tag{7}$$

The two-anomalies decision schema offers the five levels none through structure. A causal level parses and is then rejected by the validator with the message "A causal claim level is outside the declared decision scope"; rejected attempts are counted as `raw_causal_attempts`, and `unsupported_causal_claim` flags any final state at that level. When repeated repair fails to produce a valid decision object, the harness substitutes a type-neutral fallback at level none whose claim sentence is

Table 6: Claim targets, implemented evidence checks, and offline evaluation. The checks are not general identification guarantees.

Family (schema level)	Object proposed	Implemented evidence or check	Evaluation
No claim (none)	No resolved scientific conclusion	Initial or fallback state	Fallback and abstention records
Predictive (predictive)	Performance in the declared scope	Validation-based selection; private test after freezing	MSE and NMSE
Association (predictive)	Observed dependence	Observational panels and resampling	Regime coverage and schema ceiling
Variable (variable)	Candidate mechanism input	Predictor selection by refit and validation; causal membership not established	Precision, recall, F1 against synthetic parents
Functional (mechanism)	Equation or composition	Executable fit and candidate discrimination in specified regimes	Expression match and reference comparisons
Graph (structure)	Directed assignment structure	Local equations fitted to measured intermediates	Edge F1, SHD, and reported intermediate-edge score
Interventional (causal)	Response to manipulation	Outside the implemented decision scope	Attempts counted; level rejected by validator

fixed to "The current evidence does not yet distinguish the live accounts."; the initial state of every cell carries the same level and sentence.

Every typed evidence event carries a benchmark claim ceiling. In the two-anomalies cells, observational panels sampled inside the declared scope carry ceiling predictive, and the out-of-regime probe carries ceiling mechanism with its wider regime coverage; the schema therefore assigns observational-association statements to the predictive rung. Assigning the probe a mechanism ceiling encodes its designed role in this candidate-discrimination task; it does not establish a general rule for observational identification. The implemented ceiling audit requires the final declared level to stay within the strongest visible ceiling,

$$\rho(c_{\text{final}}) \;\le\; \max_{e \in \mathcal{E}_{\text{vis}}} \rho\big(\text{ceiling}(e)\big), \tag{8}$$

where $\mathcal{E}_{\text{vis}}$ is the evidence visible in the run; a violation is recorded as `claim_miscalibrated`, and the two-anomalies record reports no violation in either arm (Appendix E.6). The matched-menu appraisal of the matched-trajectory harness applies the same cap constructively: on decisive evidence it maps the diagnosed gap family to a level (variable to variable, intermediate to structure, functional family and relational program to mechanism) and releases the smaller of that level and the ceiling of the decisive panel. Passing this rule establishes only consistency with a visible level cap. It does not verify that evidence supports the specific candidate, claim, and regime, or that later evidence has not defeated that support. The historical metric names containing "calibrated" refer to this check; they denote neither statistical calibration nor a complete implementation of P5. Claim-specific evidential appraisal remains broader than the current audit.

Equivalence rules. Two proposal names denote the same quantity exactly when RESOLVECOMPONENT maps them to one canonical identifier, through normalisation, exact match, lexical overlap, and an optional semantic matcher that accepts only close rephrasings and never sees role labels. All variable-level metrics are computed after this canonicalisation, so a quantity recovered under a synonym counts as the same variable for scoring.

Two DSL strings denote the same expression when their desugared operator trees over canonical names are identical; the normalisation covers identifiers, desugaring, and whitespace and attempts no algebraic rewriting, so the exact-match flag is a conservative diagnostic beside the graph metrics of Eq. (6).

In the frozen harnesses, two hypotheses count as the same account exactly when they carry the same public candidate identifier of their sealed bundle, and candidates that agree under every admissible test collapse into one equivalence class, which the observational-equivalence cells of the diagnosis-stress comparison operationalise (Section 4.2).

A.5 EVIDENCE AND DECISION LEDGER AND SUPPORTING SAFEGUARDS

The typed evidence $\mathcal{E}_{\le t}$ and the claim record κ_t are stored as an append-only ledger of event records, and the integrity audits of Appendix D.4 run over this ledger. The safeguards around it (versioned state with provenance, storage for unresolved alternatives, typed evidence, rollback with recomputation of dependent state, group-aware removal checks, and a sealed private test with access logs) protect the

Algorithm 1 The two-level discovery loop with same-exposure propose–critique trials

Require: question q, target Y, initial variables $\mathcal{O}_0$, table A_0, corpus $\mathcal{K}_{\text{scope}}$; budgets T, J, K, S_{dsl}, patience P_{nogain}

$\mathcal{O} \leftarrow \mathcal{O}_0$; $A \leftarrow A_0$; $Q \leftarrow [\,]$
$D_0 \leftarrow \textsc{Retrieve}(q, \varnothing, \varnothing, \mathcal{O}, 0)$
$\widehat{h}_0 \leftarrow \textsc{GenerateAndValidateDSL}(q, D_0, A_0)$; $h \leftarrow \textsc{SymbolicRefit}(\widehat{h}_0, \varnothing, A_0)$
$\mathcal{H} \leftarrow \textsc{SeedHypothesisSet}(h, A_0, K)$; $h^\star \leftarrow \textsc{Best}(\mathcal{H})$; $\varepsilon_{\text{prev}} \leftarrow \text{MSE}(h^\star, A, Y)$
for $t = 1, \ldots, T$ **do**
  $D_t \leftarrow \textsc{Retrieve}(q, h^\star, \mathcal{H}, \mathcal{O}, t)$ ▷ P4
  $C_t \leftarrow \textsc{RankCandidates}(D_t, \mathcal{O}, h^\star)$
  $R \leftarrow [\,]$; $M \leftarrow \varnothing$ ▷ trials; same-exposure exclusion
  for $j = 1, \ldots, J$ **do**
    $h_j \leftarrow \mathcal{H}[((j-1) \bmod |\mathcal{H}|) + 1]$ ▷ rotate hypothesis contexts
    $\widetilde{\Gamma}_j \leftarrow \text{LLMPROPOSE}(q, D_t, h_j, \mathcal{O}, C_t, (\mathcal{O} \cup \{Y\}) \cup M)$ ▷ P2, P4
    malformed-proposal repair and premature-stop verification are internal to LLMPROPOSE
    if $\widetilde{\Gamma}_j = \text{STOP}$ **then break**
    end if
    $\Gamma_j \leftarrow \textsc{ResolveComponent}(\widetilde{\Gamma}_j, \text{aliases, semantic matcher})$
    if invalid **then** $M \leftarrow M \cup \{\widetilde{\Gamma}_j\}$; **continue**
    end if
    $V_j \leftarrow \textsc{OracleQuery}(\Gamma_j)$ ▷ values, never labels; P1
    $\widehat{\mathcal{H}}_j \leftarrow \{\textsc{GenerateAndValidateDSL}(q, D_t, A \cup V_j, h_j, \Gamma_j)\}_{1..S_{\text{dsl}}}$ ▷ P3
    $h'_j \leftarrow \arg\min_{\widehat{h} \in \widehat{\mathcal{H}}_j} \text{MSE}\big(\textsc{SymbolicRefit}(\widehat{h}, h_j, A \cup V_j, \text{required} = \Gamma_j); A \cup V_j, Y\big)$; $\varepsilon_j \leftarrow \text{MSE}(h'_j; A \cup V_j, Y)$
    $s_j \leftarrow \textsc{DecisionScore}(\Gamma_j, h'_j, \varepsilon_j, \varepsilon_{\text{prev}}, C_t)$; $R \leftarrow R + [(\Gamma_j, h'_j, s_j)]$; $M \leftarrow M \cup \{\Gamma_j\}$ ▷ P5
    early stop when consecutive non-improving trials reach P_{nogain}
  end for
  commit the best trial only if it improves validation (UPDATEHYPOTHESISSET: update A, $\mathcal{O}$, $h^\star$, $\varepsilon_{\text{prev}}$, append to Q); otherwise record no-commit and continue to the next exposure ▷ P6
end for
return $h^\star$, $\mathcal{O}$, $\mathcal{H}$, committed history Q

comparisons and add no discovery capability; ground truth, hidden-role metadata, future evidence, and private-test outcomes reach no discovery policy unless a named oracle condition grants access. Every record fixes its run position, the information visible at the event, its typed content, its parse and repair history, and its cost; exposure, retrieval, proposal, measurement, comparison, commitment, revision, and rollback events all use this format, and the released code documents the field-level schema.

Before any private row is touched, the run is fingerprinted,

$$F_{\text{run}} = \text{SHA256}\big(\text{json}(\text{cond, task, seed, } H_{\text{bundle}}, H_{\text{menu}}, H_{\text{sched}}, \kappa_{\text{dec}}, \mathcal{E}_{\text{exp}}, \mathcal{S}_{\text{fin}})\big), \tag{9}$$

the digest of a canonical serialisation of the arm, task, seed, the frozen bundle, menu, and schedule hashes, the decision objects, the exposure records, and the final state. The sealed evaluation receives only the committed candidate, the observed panel identifiers, and F_{run}, so a private outcome can influence no earlier record without changing the digest.

B PROTOTYPE DETAILS

This appendix specifies the prototype to the level of its decision points: what each subroutine may see, what it must return, and which rules admit a change of state. The measurement interface supplies values rather than relevance labels; model acceptance is governed by refitting, validation, and the declared update rules. The separate policy harnesses differ in this respect, since their appraisals are supplied.

B.1 LITERATURE RETRIEVAL AND PROGRESSIVE EXPOSURE

Retrieval is controlled and emulated: the scoped corpus $\mathcal{K}_{\text{scope}}$ is a frozen per-task document pool, and each experiment declares which subset is visible at each outer step, so a retrieval failure mode

becomes an experimental condition of the run (Section 3.2). Three objects stay separate: the corpus, the exposure policy that selects documents at a given step, and the ranking policy that scores unmeasured quantities inside the visible text.

Corpus layout and exposure schedule. The reference panel holds one frozen document pool per scenario, released over five cumulative exposure steps by a frozen schedule; step 0 is the baseline retrieval that precedes any proposal. The capability panel carries its own cumulative corpus of 144 reviewed documents (Appendix C.3).

Chunk retrieval. RETRIEVE builds a query string from the step index t, the question q, the incumbent $h^\star$, the competing hypotheses $\mathcal{H}$, and the measured set $\mathcal{O}$, then appends a retrieval objective determined by the declared strategy. Under the discriminative strategy the objective asks for evidence that distinguishes, refutes, or prunes competing hypotheses and avoids repeating support already held; under the default strategy it asks for evidence that refines, supports, or challenges the incumbent. A local GraphRAG-style adapter returns the top k evidence chunks, the extracted entities, the document identifiers, and the query itself, all entering the typed evidence and the run trace (Appendix A.5); downstream prompts carry the chunks and, when enabled, the entity hints.

Ranking policy. RANKCANDIDATES scores each unmeasured candidate from the exposed text alone, from canonical-mention, alias, and entity hits with small novelty and alias-gap bonuses; the top entries form the shortlist C_t. The score reads only discovery-visible vocabulary (the vocabulary the discovery process can see) and retrieved evidence, and hidden-role labels are excluded by construction. C_t conditions LLMPROPOSE and feeds the literature-support term of DECISIONSCORE, while commitment is settled by the validation gate of Appendix B.6.

Filter conditions and exposure state. The filter and the ranking policy vary independently: the literature-exposure suite varies the filter across nine policies with the pipeline fixed, under a declared document-budget confound, and the cumulative-evidence comparison varies the exposure state, cumulative against a delta-only override (Appendix E.3).

B.2 PROPOSAL, REPAIR, AND CANONICAL RESOLUTION

LLMPROPOSE governs one same-exposure propose–critique trial up to the point where a canonical component exists. Its inputs are the question, the exposure with its literature block, the conditioning hypothesis, the measured set, the prior requests, the same-exposure exclusions, the forbidden set, and the ranked shortlist C_t. Parsing, malformed-proposal repair, and premature-stop verification are internal to the subroutine, so the loop of Algorithm 1 receives either one raw proposal or a credible stop signal.

Prompt contract. The prompt casts the model as an iterative scientific investigator and asks for the single currently unmeasured quantity that would most improve, correct, or simplify the incumbent account. The forbidden set is assembled fresh for every trial,

$$\text{forbid}_{t,j} = \mathcal{O}_t \cup \{Y\} \cup M_{t,j}, \tag{10}$$

where $M_{t,j}$ is the same-exposure exclusion set accumulated by earlier trials of exposure t. For DeepSeek clients the prompt carries a short formatting instruction, which permits private reasoning inside a think block and forbids prose outside the requested final format, unless the task prompt already contains such a block. Complete prompt texts are reproduced in Appendix F.1.

Output schema and repair. The contract admits exactly one XML block per response: `<feature>snake_case_name</feature>` for a single quantity, `<component>name_a, name_b</component>` for a bounded component when that condition is enabled, and `<stop>I have enough information</stop>`. REPAIRPROPOSAL applies a fixed parse ladder to the raw output and, when every route fails, issues a repair prompt that carries the invalidity reason together with the original context, so a repair attempt is conditioned like the original request.

Stop verification. A stop signal is overridden when the exposed text still carries unresolved-mechanism cues or when the shortlist's top score reaches a declared threshold; the repair prompt then asks for one unmeasured underlying quantity instead, the guard is heuristic, and Section 3.3 records the corresponding boundary.

Same-exposure exclusions. M records malformed outputs, unresolvable names, requests for measured quantities, and every trialled component, so one exposure is revisited without repeats; entry carries no verdict, the exclusion clears at the next exposure, and the quantity may be proposed again under later literature.

Semantic resolution and alias rules. RESOLVECOMPONENT resolves each proposal in a fixed order: identifier normalisation, exact match, lexical overlap, an optional LLM matcher, and the alias table, with every failure recorded by reason. The matcher sees only the question, the target, the discovery-visible inventory, and evidence snippets; it accepts only close rephrasings, and a confident match is remembered as a new alias (Appendix A.4). Bounded components are enabled only in the joint-commitment study, with a maximum size of two quantities.

B.3 MEASUREMENT INTERFACE AND ORACLE SEMANTICS

ORACLEQUERY is the controlled measurement interface of the prototype. For a canonical quantity x_j and a requested sample count n it returns a value vector $v_j \in \mathbb{R}^n$ aligned with the rows already held in A. It returns numbers and no verdict: the interface never states whether x_j is a ground-truth parent, whether the request was scientifically apt, or whether the values were drawn from the simulator or from a distractor generator.

Branch structure. A repeated request returns the stored earlier values, and a first request routes by evaluation-only role metadata: essential quantities come from the simulator-backed dataset, all others from the distractor generator; both branches return a plausible series of the same length, and neither annotates it.

Distractor generation. Non-essential requests are answered with seeded independent distractor series; the interface also supports weakly correlated and correlated modes and an LLM value generator, and the reported runs use none of them.

Noise and stochasticity. Three sources of variation are kept distinct. Synthetic target values carry additive observation noise,

$$Y = f_{\mathrm{GT}}(X) + \eta_Y, \qquad \eta_Y \sim \mathcal{N}\big(0, \sigma_Y^2\big), \tag{11}$$

so a structurally correct account need not reach zero error, and the layered references of Section 4.4 inherit that floor. Proxy-specific conditions generate a proxy as a noisy function of the latent quantity it stands for,

$$x_{\mathrm{proxy}} = \alpha x + \eta_x, \qquad \eta_x \sim \mathcal{N}\big(0, (c \cdot \mathrm{sd}(x))^2\big), \tag{12}$$

with c the configured noise scale, which makes a proxy useful and imperfect. Proxy conditions are reported as a separate track, and the proxy-free main panel excludes them (Section 4.4). Distractor series form the third source, and no query-time noise is added on the fixed panels, so repeated requests reproduce exactly; the stored panel already carries the observation noise of Eq. (11) and any configured proxy noise. The executable language also admits an explicit stochastic term, and the mechanism-generation pipeline supports likelihood scoring and Monte-Carlo aggregation, while the iterative refit scores deterministic fitted accounts by mean squared error (Appendix B.5).

Absence of a correctness signal. An irrelevant request returns a series as plausible as a relevant one, so a wrong proposal cannot be detected from the response; the separating evidence arrives later, through refit, validation scoring, and the commit gate. A real oracle, a measurement, an experiment, a simulator call, or a database query, returns no correctness label either.

B.4 DSL GRAMMAR, PARSER, AND COMPILER

$\mathcal{L}_{\mathrm{DSL}}$ is a restricted executable structural-equation language. Two notions of expressivity are separated: the parser and compiler define what is executable, the refit search class of Appendix B.5 what the automated search actively considers; a valid LLM term with a product or ratio can therefore be executed and scored even though the deterministic backup basis never enumerates it.

Surface syntax and operators. The parser accepts standard infix syntax and desugars it into a canonical operator tree; a hypothesis is an assignment whose left-hand side is a variable. The operator set is small: addition, multiplication, division, power, logarithm, absolute value, exponential, sigmoid, an additive noise term, and a ternary `where` as the only conditional form. Protected forms replace undefined results with finite values, and loops, recursion, mutation, indexing, and user-defined functions are rejected, so every accepted hypothesis executes in bounded time with no side effects.

Ordered assignments. The general language represents a feed-forward structural equation model as an ordered list of assignments,

$$z_1 = g_1(X), \quad z_2 = g_2(X, z_1), \quad \ldots, \quad z_r = g_r(X, z_1, \ldots, z_{r-1}), \quad Y = f(X, z_1, \ldots, z_r), \tag{13}$$

in which every right-hand side uses input quantities or quantities defined by earlier assignments, which enforces an acyclic computation order. The default iterative condition restricts a proposed hypothesis to exactly one assignment, the flat target equation, and the validator rejects any additional assignment. The ordered-program condition lifts that restriction for measured intermediates: local equations are fitted against their measured responses and the whole program is then executed to evaluate Y, which is the executable object behind the structural recovery result of Section 4.4.

Static validation. GENERATEANDVALIDATEDSL extracts the assignment, canonicalises identifiers, and rejects a candidate that misses the target, uses an unmeasured or placeholder identifier, depends on no measured quantity, omits a newly acquired variable without explanation, exceeds the single-assignment restriction outside the ordered-program condition, or fails to parse and compile; a failing candidate is repaired against the validator errors while the retry budget lasts.

Scientific notation. Numeric literals are lexed as complete tokens. An earlier tokeniser truncated `8e-5` to `8`, inflating one scenario's private error to the order of 10^7; every affected aggregate was excluded, and all numbers in Section 4 come from the corrected parser.

B.5 HYBRID SYMBOLIC REFIT AND HYPOTHESIS MANAGEMENT

SYMBOLICREFIT turns a validated prior into a coefficient-fitted executable account. The LLM prior is neither trusted nor discarded: its additive terms are extracted and enter the candidate pool with high priority, and the same treatment is given to the conditioning hypothesis h_j of the trial, so mechanism fragments established at earlier exposures remain available. A later exposure can therefore drop an older term, replace a raw quantity by a transform, or admit a newly supported one.

Candidate-basis construction. The pool $\mathcal{B}_{\text{hyb}}$ is the union of four sources given in Eq. (14): terms extracted from the LLM prior, terms extracted from the conditioning hypothesis h_j of the trial, grammar terms covering the required and newly acquired quantities, and a budgeted backup expansion over the measured set.

$$\mathcal{B}_{\text{hyb}} = \text{Terms}(\widehat{h}) \cup \text{Terms}(h_j) \cup \text{GrammarTerms}(\Gamma) \cup \text{Expand}_{\mathcal{L}_{\text{DSL}}}(\mathcal{O}_t). \tag{14}$$

The conditioning hypothesis coincides with the incumbent $h^\star$ only when $K = 1$ or on the first rotation slot of an exposure; on the remaining slots the retained fragments come from another member of $\mathcal{H}$. Extraction strips numeric multipliers and keeps structural forms such as $\log(x)$. For each measured quantity x the grammar contributes

$$\text{GrammarTerms}(x) = \{x\} \cup \{\log(x) \text{ if } \min_i x_i > 0\} \cup \{|x|\} \cup \{|x - p| : p \in P_x\}, \tag{15}$$

where P_x is the configured pivot set, containing $p = 0$ by default and scenario anchors such as $p = 20$ when the anchor lies inside the observed range of x. Pivots are fixed by grammar and scenario configuration, so the proposal prompt never has to invent a shift value. The pool is deduplicated, and if extraction yields nothing the refitter falls back to the full grammar pool, keeping the trial executable when a prior cannot be parsed.

Coefficient fitting and selection. For a subset $B \subseteq \mathcal{B}_{\text{hyb}}$ the intercept and all coefficients are estimated jointly on training rows by ridge least squares,

$$\min_{\beta_0, \beta} \sum_i \Big(Y_i - \beta_0 - \sum_{\phi \in B} \beta_\phi \phi(A_i)\Big)^2 + \lambda_{\text{ridge}} \sum_{\phi \in B} \beta_\phi^2, \tag{16}$$

and the fitted subset is scored by a complexity-penalised criterion that trades error against term count. The search enumerates subsets up to a configured limit and keeps the best; beyond the limit it fits all pool terms at once and prunes by coefficient magnitude. The leave-one-variable-out pruning epsilon is $\epsilon_{\text{prune}} = 10^{-6}$. The retained object is the sparse additive account $h(A) = \beta_0 + \sum_{\phi \in B} \beta_\phi \phi(A)$, and for ordered programs the same procedure is applied to each local equation against its measured response before the program is executed end to end.

Required-variable semantics and provenance. A subset that represents a required quantity neither directly nor through a transform is skipped during the search, which blocks the failure mode of measuring a quantity and then silently dropping it; the trial can still lose on its score or fail the commit gate. Every pool term and selected term is tagged with its source, which is what the fixed-basis arm of Section 4.4 ablates.

Hypothesis management. SEEDHYPOTHESISSET builds at most K executable members around the refitted baseline, deduplicated by normalised DSL string and ranked; BEST returns the incumbent, and the rest serve as conditioning contexts for later trials. All members share the measured pool while each selects its own terms, so alternatives diverge until evidence reorders them. After a commit, UPDATEHYPOTHESISSET admits survivors whose quantities are all measured, deduplicates, and keeps the top K.

Budgets. Table 7 separates the three budgets that are easy to conflate. K bounds how many competing accounts survive between exposures, S_{dsl} bounds how many priors are drawn for one measured component inside a single trial, and J bounds how many propose–critique trials one exposure may run.

Table 7: Search budgets of the discovery loop, their values in the reported iterative runs, and the experimental condition that varies each. K and S_{dsl} are independent quantities.

Symbol	Value	Role	Varied by
T	5	Outer literature exposures	Retrieval and exposure studies
J	3	Same-exposure propose–critique trials	Loop-mechanism diagnostics
K	3	Competing executable accounts retained across exposures	Plural-search condition
S_{dsl}	1	DSL priors drawn per measured component within a trial	Multi-DSL condition of Section 4.4
P_{nogain}	2	Consecutive non-improving trials tolerated before an early stop	Loop-mechanism diagnostics

When $S_{\text{dsl}} > 1$, the refinement call is repeated for the same component, valid priors are deduplicated, and the best validation refit is kept.

B.6 MODEL UPDATES, REVISION, AND ROLLBACK

Three controls decide what a same-exposure trial may request and whether it changes the state, and they are independent of one another. The forbidden set of Eq. (10) restricts what may be requested, the exclusion set M prevents repetition inside one exposure, and the commit gate decides whether any trial of that exposure is admitted. Only the outer stop is terminal: the forbidden set is recomputed per trial, the exclusion set clears at the next exposure, and a refused commit leaves the same quantity proposable later.

Trial ranking and the commit gate. Within an exposure, trials are ranked by DECISIONSCORE. Let R_j be the aggregate shortlist score of the quantities in Γ_j, let $u_j = 0.05$ when the measured component appears in the refitted account and $u_j = -0.05$ otherwise, and let $g_j = -0.25$ for a generic placeholder name and $g_j = 0$ otherwise. The score is

$$s_j = \frac{\varepsilon_{\text{prev}} - \varepsilon_j}{\max(|\varepsilon_{\text{prev}}|, 1)} + 0.01 \,\text{clip}(R_j, 0, 10) + u_j + g_j, \tag{17}$$

so predictive improvement is normalised against the incumbent error scale and literature support enters as a bounded bonus. Ranking is separate from admission: the highest-ranked trial is committed only if

$$\varepsilon_{\text{prev}} - \varepsilon_\star > \delta_{\text{commit}} \qquad \text{and} \qquad \frac{\varepsilon_{\text{prev}} - \varepsilon_\star}{\max(|\varepsilon_{\text{prev}}|, \epsilon)} > \rho_{\text{commit}}, \tag{18}$$

with both margins declared per run and the gate itself disableable. A commit updates the table, the measured set, the incumbent, and the live set together; a failed gate records that no update was accepted

and retains the current model and its committed variables. The decision history can still change. This meaning of `no-commit` differs from withholding a scientific conclusion in a policy harness.

Stopping rules. The inner loop stops on a credible stop signal, on no valid candidate, or after P_{nogain} consecutive non-improving trials; the outer loop stops when the exposure budget is exhausted; both are conservative heuristics, and neither implements a formal posterior stopping rule.

Diagnostic leave-one-variable-out. After a fit, the implementation can refit with one retained quantity excluded and record the validation degradation, labelling the quantity removable under Eq. (42); the 60-cell audit it feeds is reported in Appendix E.5. The diagnostic is wired to reporting alone, and no reported discovery run deletes a quantity on its strength.

Group-aware sequential revision. The revision policy assessed in the controlled diagnostic treats declared structural groups as the unit of removal: a candidate removal must survive 120 validation resamples against a one per cent non-inferiority margin with at least 80 per cent support, at most one group is removed per action, and every attempt enters the rollback ledger. Group protection is what separates this policy from stability checking alone: without it, the same resampling still damages declared components on the constructed cases.

Protection information and its boundary. Protection metadata name the groups that may only be removed together, and the controlled diagnostic supplies them from the task construction. The evidence-gated deletion comparison replaces predictive authorisation with counterevidence eligibility for every member of a proposed group; with falsification disabled by design, the gated arm blocks all removals and loses nothing, while the predictive-only arm accepts 22 removals and loses 23 ground-truth instances (Section 4.3). Whether the loop can infer valid protection metadata from discovery-visible evidence alone remains open (Appendix G).

Declared configuration. The exclusion set, the two-level stopping rules, and the no-commit record are always on. The commit gate's margins are declared per experiment: an absolute margin of zero on the reference panel, a one per cent relative margin in the joint-commitment study, and the gate disabled in the basis-expansion, multiple-candidate, and ordered-program studies, where ranking and the stopping rules govern commitment. The semantic matcher is optional and reported per run, plateau stopping is off, leave-one-variable-out stays diagnostic-only, and grouped revision is disabled in the discovery loop, where automatic removal stays blocked unless explicit counterevidence makes an entire group eligible.

C BENCHMARKS AND SCIENTIFIC CORPORA

This appendix explains, for each task set, what makes its tasks require the capability they test. It specifies the ten reference scenarios, the eight audited capability tasks, the corpus construction and its quality gates, the data generators, and the two-anomalies construction with similar initial candidate-score patterns and a designed discriminating regime shift.

C.1 THE REFERENCE PANEL (REFERENCE TASK SET)

The reference panel supplies the first tier of Table 9 and every experiment that precedes the matched-trajectory suite in Table 11. Each of its ten scenarios is a synthetic structural causal model in a separate scientific domain. A scenario fixes a question q, a target Y, an initial measured set $\mathcal{O}_0$, a scoped corpus $\mathcal{K}_{\text{scope}}$ released under a progressive exposure schedule, and exactly two parents of Y that are hidden at the start. The hidden parents must be proposed by LLMPROPOSE, mapped onto canonical inventory entries by RESOLVECOMPONENT, and measured through ORACLEQUERY before SYMBOLICREFIT can reach the reference mechanism. Target values are generated as

$$Y = f_{\text{GT}}\big(\text{pa}(Y)\big) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\big(0, \sigma_Y^2\big), \tag{19}$$

with independent exogenous draws for every parent, so a structurally correct program still carries irreducible error and exact-match scoring is reported separately from predictive error. Table 8 lists the ten scenarios; the exogenous distributions and the per-scenario noise scales σ_Y are given in Appendix C.4.

Table 8: The ten reference-panel scenarios. Initial measured sets are the proxy-free sets used by the frozen main condition. Hidden variables are target parents withheld at the start and acquirable only through the measurement interface. Reference mechanisms are read by offline audits after a run completes and are never visible during discovery.

Scenario	Initially measured $\mathcal{O}_0$	Hidden parents	Reference mechanism (evaluation only)
Air quality PM2.5	traffic_index, humidity, billboard_color_index	industrial_emission, wind_speed	pm25 = 12 + 35*traffic_index + 40*industrial_emission + 8*humidity − 2.8*wind_speed
Battery capacity loss	charge_rate, electrolyte_purity, pack_sticker_id	cycle_count, electrode_porosity	capacity_loss = 0.003*cycle_count + 0.6*charge_rate + 7*electrode_porosity − 5*electrolyte_purity
Building energy	insulation_quality, hvac_efficiency, parking_gate_count	occupancy, outdoor_temp	energy_use = 60 + 1.1*abs(outdoor_temp − 20) + 30*occupancy − 18*insulation_quality − 22*hvac_efficiency
Concrete strength	curing_time, microvoid_fraction, mold_serial_id	aggregate_quality, water_cement_ratio	compressive_strength = 18 + 0.85*log(curing_time) − 28*water_cement_ratio + 14*aggregate_quality − 35*microvoid_fraction
Crop failure	rainfall, soil_ph, tractor_color_code	pathogen_load, root_depth	crop_loss = 20 + 0.25*abs(rainfall − 75) + 7*abs(soil_ph − 6.5) + 35*pathogen_load − 0.9*root_depth
Materials tensile strength	cooling_rate, grain_size, batch_label_code	microcrack_density, temperature	tensile_strength = 0.8*temperature + 1.2*log(cooling_rate) + 0.6*grain_size − 1.5*microcrack_density
Protein expression	nutrient_level, inducer_concentration, incubator_label_id	plasmid_copy_number, cell_stress	expression_level = 5 + 16*nutrient_level + 12*inducer_concentration + 0.22*plasmid_copy_number − 11*cell_stress
Reaction yield	catalyst_loading, impurity_level, operator_badge_id	mixing_efficiency, temperature	reaction_yield = 15 + 0.05*temperature + 22*catalyst_loading + 18*mixing_efficiency − 30*impurity_level
River daily flow	rainfall, upstream_level, algae_color_index	channel_width, flow_velocity	river_flow = 0.7*rainfall + 4.0*upstream_level + 1.3*channel_width*flow_velocity
Turbine efficiency	inlet_temp, rotor_speed, maintenance_shift_code	blade_roughness, seal_leakage	efficiency = 0.42 + 8e−5*inlet_temp + 3e−5*rotor_speed − 0.28*blade_roughness − 0.35*seal_leakage

Progressive exposure and distractor mix. A frozen schedule fixes which documents are visible at each step, so a retrieval failure mode can be named instead of dissolving into generic difficulty. The bundled corpus mixes informative papers, redundant paraphrases, same-domain rivals, near-domain distractors, and cross-domain irrelevant papers; the literature-exposure suite separates this mixture into nine declared filter policies (Appendix E.3).

Acquisition order and the proxy policy. Each scenario declares its two hidden parents as singleton reveal groups, so every acquisition is evaluated on its own; the joint-only construction needs purpose-built tasks, which the capability panel supplies. The generator also defines two noisy proxy columns and one irrelevant column per scenario; the main condition withholds the proxies from $\mathcal{O}_0$, and the proxy-contamination track is the separate condition that exposes them (Appendix H.1).

Role of the panel. Each scenario runs with three prespecified seeds and, in paired designs, two conditions, giving 60 cells and 30 pairs; the panel carries the split-integrity check, the streaming, layered-reference, exposure, and cumulative-evidence comparisons, the numeric-preview ablation, the deletion-safety audit, and the proxy track. Its mechanisms are additive in a small set of univariate terms, and an open-ended basis-function regression can already express nine of the ten, so the basis-expansion and structural-recovery comparisons instead use the capability panel, where these restrictions are relevant. Ground-truth lists and reference equations are read only by offline audits after a run completes.

C.2 The capability panel (capability task set)

The capability panel contains eight purpose-built, proxy-free tasks. Domain vocabulary, variable semantics, units, and plausible operating ranges are reused from the reference-panel corpus, while every causal structure is new. Each task is designed to expose a specified limitation of a simpler control. Numerical audits check selected contrasts under the configured estimator and acceptance rule; they do not prove that all algorithms lacking a broadly defined capability must fail (Section 4.1). The eight families are interaction-only, joint-only, threshold, piecewise, nested-ratio, competing-DSL, explicit-intermediate, and multi-hop mechanisms; the full reference programs, ordered assignments in the executable language together with the matched control each is built to defeat, ship with the released benchmark. The cumulative literature for the panel contains 144 independently reviewed documents and 115,475 words, constructed as described in Appendix C.3.

The numerical design audit (called the necessity audit in the experimental record) checks the joint-only and threshold contrasts before any discovery run. Its conclusion is relative to the configured refit procedure and improvement threshold. For an incumbent program h over the currently measured variables and a proposed addition Γ, the audit computes the relative validation improvement

$$\Delta_{\mathrm{val}}(\Gamma) = \frac{\mathrm{MSE}_{\mathrm{val}}(h) - \mathrm{MSE}_{\mathrm{val}}(h \oplus \Gamma)}{\mathrm{MSE}_{\mathrm{val}}(h)}, \tag{20}$$

where $h \oplus \Gamma$ is the program refitted after Γ is admitted. A joint-only task is accepted only when Δ_{val} is negligible for each singleton and large for the pair. On the materials interaction task either singleton moves validation error by less than 0.01% while the joint interaction improves it by approximately 98.0%; on the river conveyance task the singleton bound is 0.02% and the joint improvement approximately 98.8%. On the building threshold task the piecewise intermediate program improves validation error by approximately 99.3% over the global absolute-temperature control. On the battery competing-laws task the audit confirms the intended candidate ordering, with the accelerated-threshold program ahead of the quadratic rival and the forced-first linear program last.

Data policy and split roles. Every capability task draws 256 training, 256 validation, and 512 sealed private-test rows, with the usual roles and one-time release (Appendix D.1). One task uses the declared allowance for a discriminating validation regime: the battery competing-laws task stratifies training toward mild operation while validation covers the full range, so the forced-first linear candidate is competitive in training and defeated in validation.

Task-to-comparison assignment. The basis-expansion study uses five tasks testing limitations of the predefined basis library; the multiple-candidate study uses the battery task, the only one with a forced-first plausible-but-wrong candidate; the joint-commitment study uses the two joint-only tasks;

and the ordered-program study uses the two multi-hop tasks with fresh confirmatory seeds. Each task ships a machine-readable ground-truth package, readable by evaluators after a model is frozen and by no component of the discovery loop.

C.3 Literature construction and quality control

Both corpora are synthetic scientific literature written for a declared task specification. Neither claims to reproduce real publications, and neither may contain a bibliographic citation, because a fabricated reference would be both scientifically misleading and a covert channel for authority that the discovery loop cannot check.

Source reuse and scope. The reference corpus is released under a five-step cumulative exposure schedule that fixes which documents are visible at each step of each scenario. The capability corpus reuses it as a source of domain vocabulary, variable semantics, plausible operating ranges, and reusable mechanism fragments, and writes an independent corpus for the eight new causal structures. Reuse is confined to scientific surface material and reusable mechanism fragments: every capability task defines its own acyclic SCM, and no capability task inherits a reference-panel causal structure.

Generation and review procedure. Each capability task receives 18 documents on a fixed plan of genres and evidential purposes, from cross-laboratory baselines through rival mechanisms, negative controls, and skeptical reviews, released over five cumulative exposures. Documents were drafted against the private task specification and every batch was reviewed by an independent call; the audit records 144 approved documents from 43 draft, 71 review, and 30 revision calls, with batches rejected for outcome leakage, invented mechanisms, and incorrect edge or regime descriptions.

Length, diversity, and adversarial content. The realised corpus holds 144 documents and 115,475 words, a mean of 801.9 words per document, no fabricated citations, and no near-duplicate pair. Every task carries a credible rival mechanism, a negative control, a boundary-regime conflict, and a skeptical review, and hidden quantities appear under aliases so resolution has real work to do.

Deterministic corpus checks. A deterministic validator, independent of the generator, enforces the plan: presence and exposure position, length bounds, no proxy terminology, no bibliographic patterns, no verbatim ground-truth program, pairwise token-overlap ceilings, and agreement between each stored SCM and the frozen specification; the frozen report records 8 of 8 tasks and 144 of 144 documents passing.

Frozen provenance record. The benchmark was frozen on 2026-08-02 with digests binding every released document to the specification and quality report it was written against, so a document edited after review is detectable from the record alone; the record also confirms that no credential was written into any artifact, and the digest listing ships with the reproducibility material (Appendix F.4).

C.4 Data generation, stochasticity, and distractors

This subsection separates four sources of uncertainty that are easy to conflate: observation noise on the synthetic target, the values returned for non-essential proposals, the resampling semantics of the streaming condition, and the regime structure of the two-anomalies generators.

Structural models and exogenous draws. Every reference-panel parent is drawn independently from a declared distribution, the target follows Eq. (19), and the generator is deterministic given a scenario and a seed, with every panel row-aligned across all columns including the still-unmeasured ones. The per-scenario noise scales span more than two orders of magnitude, which is why cross-task means of raw MSE are always reported beside paired directions, medians, and scale-normalised errors; the full distribution and noise tables ship with the released benchmark.

Plausible non-essential measurements. A proxy column is a noisy linear readout of one hidden parent under Eq. (12), with $\alpha \in [0.76, 0.88]$ and $c \in [0.16, 0.25]$ across the panel's twenty proxies, so each proxy is useful and imperfect at once. Irrelevant columns are generated from their own names so they stay domain-plausible, and OracleQuery returns stored simulator values for essential requests

and seeded distractor series otherwise, in both cases numbers with no label. No query-time noise is added on the fixed panels, so repeated requests reproduce exactly.

Panel construction and acquisition. A panel holds 128 row-aligned samples split 77/26/25 by a seed-derived deterministic partition; acquiring a variable unlocks its aligned column while rows, split boundaries, and target values stay fixed, which is what makes a paired comparison exact.

Streaming semantics. The streaming condition draws a fresh 128-row panel from the same SCM before each outer exposure, deterministically from the task and step, while acquired names, hypotheses, and the incumbent persist; it is observational resampling from a fixed SCM and carries no interventional content (Section 4.2).

Two-anomalies generators. The two-anomalies suite uses 192 training rows and 192 validation rows per stage and a sealed 512-row private test drawn half from the declared operating scope and half from the wider regime. Each task exposes a recorded coordinate, an auxiliary readout z, and a target y. In the two N-type tasks the auxiliary readout is a genuine input:

$$z = \gamma|\tilde{x}| + \nu, \qquad y = \beta_1 x + \beta_2 z + \varepsilon, \tag{21}$$

with $x \sim \mathcal{U}(-1.4, 1.4)$ inside the declared scope and $x \sim \mathcal{U}(-3.0, 3.0)$ outside it, and with $\tilde{x} = x$ inside the scope and $\tilde{x} \sim \mathcal{U}(-1.4, 1.4)$ drawn independently outside it. The alloy latent-parent task uses $(\gamma, \beta_1, \beta_2) = (0.975, 0.55, 1.15)$ with $\nu \sim \mathcal{N}(0, 0.085^2)$ and $\varepsilon \sim \mathcal{N}(0, 0.14^2)$, and the bioprocess latent-parent task uses $(0.945, 0.45, 1.05)$ with $\nu \sim \mathcal{N}(0, 0.09^2)$ and $\varepsilon \sim \mathcal{N}(0, 0.13^2)$.

In the V-type task with a piecewise thermal response the target depends on the recorded coordinate alone through a piecewise response,

$$g(x) = -0.45\min(x, 0) + 1.40\max(x, 0), \qquad y = 0.25 + g(x) + \varepsilon, \tag{22}$$

with $\varepsilon \sim \mathcal{N}(0, 0.13^2)$ and the same in-scope and out-of-scope ranges for x. Its auxiliary readout tracks the same signal inside the scope and is regenerated from an independent in-scope draw x' outside it,

$$z = \begin{cases} g(x) + 0.26\,\varepsilon + \zeta, & \text{inside the declared scope,} \\ g(x') + \zeta, & \text{outside it,} \end{cases} \tag{23}$$

with $\zeta \sim \mathcal{N}(0, 0.05^2)$ and $x' \sim \mathcal{U}(-1.4, 1.4)$. The second V-type task composes a drive x and a resistance r into a nested transport ratio,

$$s = \frac{x}{1 + |r|}, \qquad y = 0.2 + 1.6\,s + \varepsilon, \tag{24}$$

with $x \sim \mathcal{U}(0.25, 2.25)$, $\varepsilon \sim \mathcal{N}(0, 0.085^2)$, and $r \sim \mathcal{U}(-1.05, 1.05)$ inside the declared scope against $r \sim \mathcal{U}(-2.4, 2.4)$ outside it. Its auxiliary readout is $z = s + 0.43\,\varepsilon/1.6 + \zeta$ inside the scope and $z = x'/(1+|r'|) + \zeta$ outside it, with $\zeta \sim \mathcal{N}(0, 0.03^2)$, $x' \sim \mathcal{U}(0.25, 2.25)$, and $r' \sim \mathcal{U}(-1.05, 1.05)$.

The two families share one out-of-scope operation: the draw that severs the dependence of z on the recorded coordinate. In the N family the target still depends on the realised z, so an account that retains the auxiliary readout transfers to the wider regime. In the V family the target depends on the recorded coordinate alone, so the same account collapses out of scope while the correct structural candidate transfers. This is the registered single source of family discrimination, and it is visible only through the out-of-regime probe.

Frozen generator-level calibration. The generator design aims to make initial candidate-score patterns similar across the two families; its calibration is fixed before each run. For a candidate program c and the base model over the recorded inputs, the calibration quantity is the relative error reduction

$$\Delta_{\mathrm{rel}}(c) = \frac{\mathrm{NMSE}(h_{\mathrm{base}}) - \mathrm{NMSE}(h_c)}{\mathrm{NMSE}(h_{\mathrm{base}})}, \tag{25}$$

computed on 100k-row populations matched to the proposal-time panel distribution and never on run panels, so that no sample-level gate can be used to select seeds. On all four tasks the auxiliary candidate falls in $[0.88, 0.98]$ and both structural candidates fall in $[0.78, 0.95]$, and the auxiliary candidate is marginally the best in-scope predictor. The in-scope NMSE triples, normalised to the best candidate,

are shape-matched across families, at $(1, 1.494, 2.152)$ for alloy, $(1, 1.534, 2.144)$ for bioprocess, $(1, 1.442, 2.176)$ for thermal, and $(1, 1.556, 2.262)$ for transport. Out of scope, every wrong candidate is separated from the correct one by a margin of at least 0.9669, against a registered minimum of 0.25. A leave-one-out nearest-centroid classifier over shape-normalised triples from 200 fresh 192-row proposal panels per task separates N from V at accuracy 0.601, inside the prespecified acceptance band $[0.35, 0.65]$. This limited classifier check does not establish chance-level discrimination or statistical indistinguishability of the families. The calibration record is hashed, the hash is verified at every launch, and the values are re-derived from the generators by a deterministic test.

C.5 Matched-trajectory tasks and generators

The matched-trajectory suite supplies the matched policy cells of Section 4.2. Its five frozen comparisons, matched evidence, reopening, matched menu, diagnosis stress, and component revision, contribute 108 cells and 54 matched pairs over twelve tasks. Four shared repair instances (a missing parent, an unmeasured intermediate, a functional-family gap, and a relational-program gap) serve the matched-evidence, matched-menu, and reopening comparisons; four scripted appraisal panels serve diagnosis stress; and two removal with two retention tasks serve component revision. Induced limitations are the generators' offline labels, withheld from the policy and the validator and read only after the controller output is frozen. This subsection specifies what every generator samples, where its branch points sit, and how the two arms of a pair are held identical; the per-comparison endpoints, gate tables, and trajectory records are reported in Appendices E.2–E.5.

Repair generators shared by three comparisons. One sampler produces the four repair instances, and a single stage index k governs every branch in it. Stages $k \in \{0, 1, 2\}$ carry the three typed evidence events, and the sealed private panel is drawn last at $k = 3$, so it takes the same branch as the decisive stage. In the alloy missing-parent task the target has two parents,

$$x \sim \mathcal{U}(-1.5, 1.5), \quad z \sim \mathcal{N}(0, 1), \quad y = 0.9\,x + 1.7\,z + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 0.16^2), \tag{26}$$

and the branch drops the z column from the released panel for $k < 2$, so the second parent becomes measurable only at the targeted-measurement stage. In the reactor unmeasured-intermediate task an intermediate stands between the inputs and the target,

$$m = x\,u + \nu, \qquad y = 0.35\,x + 1.55\,m + \varepsilon, \tag{27}$$

with $x \sim \mathcal{U}(-1.2, 1.2)$, $u \sim \mathcal{U}(0.4, 1.6)$, $\nu \sim \mathcal{N}(0, 0.025^2)$, and $\varepsilon \sim \mathcal{N}(0, 0.11^2)$; the m column is withheld for $k < 2$ on the same branch. The two hypothesis-language instances keep every column measured throughout and move the sampling range instead. In the thermal functional-repair task,

$$g(x) = -0.45\min(x, 0) + 1.75\max(x, 0), \qquad y = 0.3 + g(x) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 0.09^2), \tag{28}$$

with $x \sim \mathcal{U}(-a_k, a_k)$ and $a_k = 1.0$ for $k < 2$ against $a_k = 2.3$ for $k \geq 2$, so both slopes occur in both regimes, while the wider range increases their signal magnitude. The auxiliary decorrelation below supplies an additional source of separation. In the transport relational-repair task,

$$s = \frac{x}{1 + |r|}, \qquad y = 0.2 + 1.6\,s + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 0.06^2), \tag{29}$$

with $x \sim \mathcal{U}(0.25, 2.25)$ and $r \sim \mathcal{U}(-b_k, b_k)$ for $b_k = 0.8$ at $k < 2$ against $b_k = 2.0$ at $k \geq 2$. Both language instances also expose an auxiliary readout whose dependence on the recorded coordinate is severed at the same branch,

$$z_k = \begin{cases} w, & k < 2, \\ \eta_k, & k \geq 2, \end{cases} \tag{30}$$

where w is the task signal, $g(x)$ for thermal and s for transport, and the replacement draw is $\eta_k \sim \mathcal{N}(0, 1)$ for thermal against $\eta_k \sim \mathcal{N}(0, 0.8^2)$ for transport. Inside the first regime the auxiliary readout equals the signal exactly, which is what leaves the auxiliary account and the structural account observationally indistinguishable before the decisive event. The three comparisons read this one trace differently. The matched-evidence comparison replays all three stages in order. The reopening comparison releases stage 0 as the provisional outcome and delivers stages 1 and 2 as the non-decisive and decisive monitoring events. The matched-menu comparison seals the stage outcomes behind its action menu, mapping stage 2 to the discriminating action, stage 1 to the same-scope replication, and stage 0, relabelled as a precision audit, to the generic action.

Component-revision generators. The component-revision sampler has one regime index with two values, `initial` for the two panels that form the current account and `later` for the declared new observational regime and the sealed private rows. All four tasks share an additive target noise $\varepsilon \sim \mathcal{N}(0, 0.12^2)$ and a floor of 32 rows per panel. The two removal tasks place a component in the account that the initial regime supports and the later regime does not,

$$y = 2.0\,p + \varepsilon, \qquad y = 1.4\,p - 0.8\,c + 0.9\,pc\,\mathbf{1}[\text{initial}] + \varepsilon, \tag{31}$$

with $p, c \sim \mathcal{N}(0, 1)$. In the first task the challenged auxiliary column is $a = p + \eta$ with $\eta \sim \mathcal{N}(0, 0.04^2)$ in the initial regime and $a \sim \mathcal{N}(0, 1)$ in the later regime, so the auxiliary term is collinear with the true parent at first and carries nothing afterwards. In the second, the interaction is present only in the initial regime. This is a regime-dependent change in the generating equation and more than an acquisition artifact; removing it is appropriate to the later regime. The two retention tasks invert that arrangement,

$$y = 1.1\,p + 1.3\,q + \varepsilon, \qquad y = 2.0\,o + 0.9\,\max(o, 0) + \varepsilon, \tag{32}$$

where $q = p + \eta$ with $\eta \sim \mathcal{N}(0, 0.05^2)$ in the initial regime and $q \sim \mathcal{N}(0, 1)$ in the later regime, and where $o \sim \mathcal{U}(-2.0, -0.05)$ in the initial regime against $o \sim \mathcal{U}(-2.0, 2.0)$ in the later one. The challenged components are a genuine input in the first task and a genuine nonlinear term in the second. Their apparent redundancy depends on the initial sampling regime.

Sealed panels, shared indices, and counterbalancing. The repair comparisons draw 192 training and 192 validation rows per stage with a sealed 512-row private panel, and the component-revision comparison 96/96 with 192 sealed; every panel is rebuilt deterministically from the task identifier and seed, the sealed rows share no draw with any stage panel, and private scoring refits the selected candidate on the final stage's training columns. Both cells of a pair therefore observe identical rows, bundles, and evidence traces, which the matched-input hash gate verifies, and candidate and menu order are counterbalanced by seed-derived parity with neutral public identifiers. A generator-side audit gates every instance before any cell runs, checking answer-term leakage, the ambiguity margin of the early events, and the decisiveness of the final event, and private release is bound to a fingerprint of the frozen final state through Eq. (43).

The diagnosis-stress comparison as a scripted appraisal panel. This is the one comparison with no SCM sampler: each instance is a scripted appraisal panel with hand-set per-candidate scores and zero training, validation, and private rows, its only stochastic element a small seeded presentation jitter. Nothing predictive can be selected, stopped, or tuned inside it, so private testing is declared inapplicable; the estimand is the joint diagnosis-and-response under frozen evidence, and its pairs enter no private-test aggregate in this paper.

D Evaluation protocol

This appendix records the protocol behind every reported number: the row roles and the one-time release of the private test, the metric definitions, the seed and gate registers, and the audits that abort a run instead of letting it drift.

Table 9 summarises the four evidence tiers and their roles; the subsections below fix the protocol they share.

D.1 Train, validation, and private-test roles

Every number in Section 4 is scored on rows whose role was fixed before the run began; three roles are used throughout. Training rows fit coefficients and select the sparse basis subset inside SymbolicRefit, and they supply the numeric preview shown to LLMPropose and GenerateAndValidateDSL. Validation rows compare same-exposure propose–critique trials through DecisionScore, decide commit and no-commit, govern stopping, and authorise revision. Private-test rows measure the frozen account once, after the discovery state has been fingerprinted. The three index sets are pairwise disjoint and are shared by every column of a panel, so a variable acquired through OracleQuery extends the value table A_t sideways while the row partition stays fixed.

Table 9: Evidence tiers and their scientific roles. Design audits test limitations of specified controls; conclusions are restricted to the reported configurations and tasks.

Tier	Scientific content	Role	Scale
Reference panel	Mostly flat additive mechanisms	Pipeline validity, observational streaming, layered reference errors, evidence-state sensitivity	10 scenarios; 60 cells; 30 pairs
Capability panel	Interaction, joint-only, threshold, piecewise, ratio, competing hypothesis language, intermediate, multi-hop mechanisms	Necessity-audited representation and abduction tasks	8 tasks; 144 documents
Matched trajectories	Controlled missing-variable, missing-intermediate, functional-family, relational-program, equivalence, and measurement-reliability cases	Policy-isolated trajectories: evidence choice, typed response, reopening, no-commit, safe revision	5 comparisons; 108 cells; 54 pairs
Two anomalies	Matched task families in which an auxiliary readout is a genuine input or merely tracks the target	Retention-direction comparison under identical menus and a shared exhausted budget	4 tasks; 24 main and 6 forced-probe cells; 12 pairs

Row partition of the prototype panels. A reference panel holds $n = 128$ row-aligned observations, shuffled with a seed-derived split seed and cut by frozen fractions into 77 training, 26 validation, and 25 private-test rows; the index lists are stored with the run, and every column, including newly acquired ones, is indexed by the same three sets. In the streaming regime a fresh panel is drawn before each exposure under the same rule. Training rows never enter candidate comparison, stopping, or reporting; validation rows never enter coefficient fitting; private rows enter nothing until the one-time release.

Row budgets in the frozen policy harnesses. The matched-evidence, matched-menu, reopening, and two-anomalies comparisons record 192 training and 192 validation rows per stage with a 512-row private test; component revision uses 96/96/192; the diagnosis-stress comparison declares all three roles empty and private testing inapplicable. The two-anomalies private panel is mixed-regime, half inside the declared scope and half from the wider regime, so the sealed measurement covers where an in-scope predictor and a structural account come apart. The real-data analogue splits its 569-row table into 341/114/114 three times.

Decision freezing and the one-time release. The private panel is physically absent from every prompt, appraisal, and analysis input while discovery runs; release requires the fingerprint of the frozen trajectory (Eq. (9)), a second fingerprint is refused, and any later edit of the trajectory invalidates the release. A cell that ends with no commitment records null private values.

End-to-end check of the split path. The split-integrity check exercises the three roles end to end on one panel: private-test MSE falls from 23789.829 to 0.253 on rows that played no part in fitting or selection, and the final errors order 0.135, 0.173, and 0.253 across the roles; a third-step no-commit leaves all three role curves bit-identical. One scenario and one split seed make this a protocol smoke test and nothing more.

Declared exception. The real-data feature-selection analogue computes and stores private-test metrics for every candidate trial in its step trace, although those values enter no automated decision and its progressive commit rule reads validation MSE alone at a permissive threshold of 10^{-6}. This falls short of the release-once rule above, and Section 4.5 reports that result as a boundary analogue on that ground.

D.2 Metrics and claim-level scoring

The implementation records metrics at three granularities: a per-session summary, per-step diagnostics, and per-trial records; the frozen policy harnesses add per-cell metrics and paired deltas. Throughout this subsection, $h^\star$ is the frozen final executable account, $\mathrm{vars}(h^\star)$ its canonical non-target variables after ResolveComponent, $\mathrm{pa}(Y)$ the evaluation-only ground-truth parent set of the target, H the ground-truth mechanism variables that are unmeasured in $\mathcal{O}_0$, D_{final} the acquired identifiers retained in $h^\star$, and Q the ordered history of unique requested and committed variables. All variable-level quantities are computed after canonical resolution, so a quantity recovered under a synonym counts as the same variable claim (Appendix A.4).

Variable recovery on the final account. The primary variable metrics are computed on the frozen account, so a variable that was proposed and then dropped by refit or by the no-commit rule leaves them unchanged, while a variable that was proposed and retained enters them:

$$P_{\mathrm{var}} = \frac{|\mathrm{vars}(h^\star) \cap \mathrm{pa}(Y)|}{|\mathrm{vars}(h^\star)|}, \qquad R_{\mathrm{var}} = \frac{|\mathrm{vars}(h^\star) \cap \mathrm{pa}(Y)|}{|\mathrm{pa}(Y)|}, \qquad F1_{\mathrm{var}} = \frac{2P_{\mathrm{var}}R_{\mathrm{var}}}{P_{\mathrm{var}} + R_{\mathrm{var}}}. \quad (33)$$

Hidden-variable recovery. The hidden view restricts attention to the variables that discovery had to acquire. Its recall denominator is H, and its precision denominator is D_{final}, the subset of final-account variables that are hidden benchmark variables or were requested during discovery:

$$P_{\mathrm{hid}} = \frac{|D_{\mathrm{final}} \cap H|}{|D_{\mathrm{final}}|}, \qquad R_{\mathrm{hid}} = \frac{|D_{\mathrm{final}} \cap H|}{|H|}, \qquad F1_{\mathrm{hid}} = \frac{2P_{\mathrm{hid}}R_{\mathrm{hid}}}{P_{\mathrm{hid}} + R_{\mathrm{hid}}}, \quad (34)$$

Eq. (44) instantiates the same quantities for the literature-exposure suite; Eq. (33) is the cleaner precision statement against all target parents, Eq. (34) isolates the quantities obtained through inquiry, and the graph metrics below carry the structural claim.

Functional equivalence and hidden coverage. Let ν be the normalisation of Appendix A.4, covering identifier canonicalisation, desugaring to the canonical operator tree, and whitespace, and let h^{gt} be the benchmark functional form. The strict equivalence flag and the coverage of the hidden mechanism variables are

$$m(h^\star) = \mathbf{1}\big[\nu(h^\star) = \nu(h^{\mathrm{gt}})\big], \qquad c(h^\star) = \frac{|\mathrm{vars}(h^\star) \cap H|}{|H|}, \quad (35)$$

recorded as the exact-match flag and the hidden-coverage fraction of the frozen account, with a stepwise form of c recorded at every outer step. Because ν attempts no algebraic rewriting, an algebraically equal reformulation scores $m = 0$, and m is read as a conservative diagnostic beside the graph metrics. An earlier weighted hypothesis-accuracy heuristic is retired from the main interpretation, since it mixed target definedness, hidden coverage, and exact string match without measuring graph structure.

Target-edge and full-graph structure. A flat target equation induces the predicted and benchmark target-edge sets of Eq. (6), scored by the edge precision, recall, and F1 of Eq. (48) and the structural Hamming distance of Eq. (49), together with the counts of false-positive and false-negative edges. An ordered program with local equations extends the comparison to the full assignment graph. Writing $h^\star$ as a set of assignments $Z \leftarrow f_Z$ over intermediates and the target,

$$\widehat{E}_{\mathrm{full}} = \bigcup_{(Z \leftarrow f_Z) \in h^\star} \big\{(x, Z) : x \in \mathrm{vars}(f_Z)\big\}, \quad (36)$$

and the same three edge scores and the same Hamming distance are computed against the benchmark edge set $E^\star_{\mathrm{full}}$, with reversed ground-truth edges counted separately. Intermediate-edge recall restricts recall to the edges whose head is a measured intermediate,

$$R_{\mathrm{int}} = \frac{|\widehat{E}_{\mathrm{full}} \cap E^\star_{\mathrm{int}}|}{|E^\star_{\mathrm{int}}|}, \qquad E^\star_{\mathrm{int}} = \{(x, Z) \in E^\star_{\mathrm{full}} : Z \neq Y\}, \quad (37)$$

which would be zero for a strictly flat target equation under this definition. Table 22 nevertheless reports nonzero flat-arm values (0.250 and 0.400). We preserve those recorded values, but the implementation of this endpoint must be reconciled with the stated definition; the main structural comparison therefore relies on full-graph F1 and SHD.

Prediction and executable utility. For an account h and an index set $\mathcal{I}$ the two error scales are

$$\mathrm{MSE}(h; \mathcal{I}) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \big(Y_i - h(X_i)\big)^2, \qquad \mathrm{NMSE}(h; \mathcal{I}) = \frac{\mathrm{MSE}(h; \mathcal{I})}{\max\big(\widehat{\mathrm{Var}}\,(Y_{\mathcal{I}}),\ 10^{-12}\big)}, \quad (38)$$

with $\widehat{\mathrm{Var}}\,(Y_{\mathcal{I}})$ the variance of the target on the scored rows. NMSE is the harness-level scale and makes cells with different target magnitudes comparable. With $\varepsilon_{\mathrm{base}}$ the error of the baseline account fitted from $\mathcal{O}_0$ and the baseline exposure, $\varepsilon_{\mathrm{final}}$ the error of $h^\star$, and ε_t the error after the committed state of outer step t,

$$\Delta_{\mathrm{final}} = \varepsilon_{\mathrm{base}} - \varepsilon_{\mathrm{final}}, \qquad \Delta_t = \varepsilon_{\mathrm{base}} - \varepsilon_t, \qquad \delta_t = \varepsilon_{t-1} - \varepsilon_t, \quad (39)$$

so positive values favour the later state. The trial-level form of δ, with $\varepsilon_{\text{prev}}$ and ε_j as in Algorithm 1, is the same-exposure improvement signal read by the inner loop and by the no-commit rule. A model-agnostic companion metric scores the available feature set with a deterministic nonlinear evaluator, separating variable usefulness from functional-form correctness.

Proposal process and query efficiency. The search process is scored over the request history Q, which includes over-discovery that the final account later dropped:

$$P_{\text{prop}} = \frac{|Q \cap H|}{|Q|}, \qquad R_{\text{prop}} = \frac{|Q \cap H|}{|H|}, \qquad F1_{\text{prop}} = \frac{2P_{\text{prop}}R_{\text{prop}}}{P_{\text{prop}} + R_{\text{prop}}}. \tag{40}$$

Query efficiency counts relevant unique proposals against unique proposals,

$$\eta = \frac{|Q \cap H|}{\max(1, |Q|)}, \tag{41}$$

which is numerically the same family of diagnostic as P_{prop}.

Trajectory and inner-loop diagnostics. Order-sensitive diagnostics over the request history, proposal-order average precision, cumulative hidden recall, revisit benefit, commit regret, and hypothesis-set diversity, are recorded per run; their definitions ship with the released code.

Revision safety. Deletion diagnostics are scored on validation rows. For a variable x retained by $h^\star$, let $h^\star_{-x}$ be the account refitted without it. The leave-one-variable-out degradation and the removability flag are

$$d(x) = \text{MSE}\big(h^\star_{-x}; \mathcal{I}_{\text{sel}}\big) - \text{MSE}\big(h^\star; \mathcal{I}_{\text{sel}}\big), \qquad \text{rem}(x) = \mathbf{1}\big[d(x) \le \epsilon_{\text{prune}}\big], \tag{42}$$

with the frozen $\epsilon_{\text{prune}} = 10^{-6}$. An exclusion with $\text{rem}(x) = 1$ and $x \in \text{pa}(Y)$ is a dangerous prune, and the ground-truth-variable loss count of a revision arm is the number of accepted deletions whose removed variable lies in $\text{pa}(Y)$; the ground-truth edge loss is its edge-level analogue over Eq. (6). The component-revision endpoints are fractions over matched pairs: useful-removal recall counts the pre-specified components that fail the frozen across-resample transfer test and are removed, warranted retention counts the challenged components that pass that test and are kept, and coherent revision requires the superseded claim to be withdrawn and dependent state recomputed alongside the removal.

Claim-level scoring. Every released decision carries a level on the ladder of Eq. (7) and is audited against the strongest visible ceiling by Eq. (8); violations flag the cell, causal attempts are counted separately, and per-arm fallback rates are the denominator correction applied before control-arm gates are evaluated.

D.3 Matched budgets and statistical summaries

Main comparisons use DeepSeek-R1-Distill-Qwen-32B with three prespecified seeds per task instance, and every reported contrast is paired within a task and seed (Section 4.1). A pair shares the task generator, the data rows, the available literature or evidence, the initial candidates, the model profile, the budgets, and the evaluation code, so the arms differ only in the declared treatment. This subsection records the seed lists, the cell and pair counts, the matched budgets, and the summaries used to aggregate them.

Model choice and transferability. All main confirmatory runs use DeepSeek-R1-Distill-Qwen-32B with a pinned checkpoint and temperature 0. This supports reproducibility without reliance on a changing commercial endpoint; exact numerical determinism also depends on the recorded runtime and hardware. To check the evidence-selection result under another model, a small transferability study reruns the matched-menu comparison, the cheapest and most tightly controlled of the five matched-trajectory comparisons, with a frontier commercial model (GPT-5.6) behind the same harness, prompts, menus, seeds, and gates, at the endpoint's default decoding temperature, the one setting it does not let a caller pin; Table 10 reports both models side by side. The claim under test is directional: the uncertainty-resolving selector should improve evidence efficiency under either model, so the study checks trend agreement and does not compare absolute scores across models. The rerun agrees in

Table 10: Model transferability on the matched-menu comparison: the frozen main result (DeepSeek-R1-Distill-Qwen-32B) against a rerun with GPT-5.6 under identical harness, prompts, menus, seeds, and gates.

Gate or metric (12 pairs)	DeepSeek-R1-32B	GPT-5.6
First discriminating action, uncertainty-resolving arm	12/12	12/12
First discriminating action, support-seeking arm	9/12	2/12
Mean actions to decisive evidence, uncertainty-resolving arm	1.00	1.00
Mean actions to decisive evidence, support-seeking arm	1.25	2.08
Correct final account, uncertainty-resolving arm	12/12	12/12
Correct final account, support-seeking arm	12/12	9/12
Private-test MSE noninferior within pair (5% margin)	12/12	12/12
Prespecified acceptance decision	PASS	PASS

direction on every gate and passes the same prespecified acceptance decision; under GPT-5.6 the support-seeking baseline is weaker than under DeepSeek, which enlarges the efficiency difference. Unlike the main-model comparison, final correctness also differs (12/12 versus 9/12); this study does not test transfer of the other revision policies. The rerun is archived with the same manifest and analysis layout as the main comparison (run-20260906-182444).

Seeds and their lineage. The prototype tiers use evaluation seeds 0, 1, and 2, except for the confirmatory ordered-program matrix, which uses fresh seeds 12, 13, and 14 after four earlier iterations on seeds 0 to 2, 3 to 5, 6 to 8, and 9 to 11 that are retained as development lineage. Each frozen matched-trajectory comparison draws a fresh seed triple: 1301, 1409, and 1511 for matched evidence; 401, 503, and 607 for the matched menu; 1601, 1709, and 1811 for reopening; 3301, 3407, and 3511 for diagnosis stress; and 2503, 2609, and 2711 for component revision. Every superseded formulation holds its own fresh triple, and its rates are never pooled with the accepted comparisons. The two-anomalies suite uses 4201, 4303, and 4409 in its first run, 5101, 5209, and 5303 in the repaired run, and 6101, 6211, and 6307 in the confirmatory run. A seed-collision audit over every frozen design records no reuse of a seed triple across designs, which keeps a repaired design from inheriting the cells that motivated the repair.

Paired cells. The reference panel contributes 60 cells and 30 pairs, the capability studies 30, 6, 12, and 12 cells, the exposure suite 108 sessions, the cumulative comparison 240 cells and 120 pairs, the five matched-trajectory blocks 108 cells and 54 pairs, and the two-anomalies suite 30 cells, of which 24 form 12 matched pairs and 6 are the single-arm forced-probe component. The budgets each comparison holds fixed inside a pair are declared in its frozen design; Table 29 lists the six confirmatory harness designs.

Literature, token, and measurement budgets. The frozen prototype configuration exposes at most 10 chunks per exposure, ranks a shortlist of 5, runs at most 5 exposures with 3 same-exposure trials, keeps 3 accounts, and decodes at temperature 0. Generation caps are part of the measured design: 1024 new tokens for the prototype tiers and most harness comparisons, 2048 for diagnosis stress, and 4096 for the repaired and confirmatory two-anomalies runs, after the first run failed its comparative gate under the 1024-token cap. Harness budgets are counted in equal-cost evidence actions, two in the matched-menu comparison and three in the two-anomalies comparison, with at most two parse attempts per decision before the type-neutral fallback.

Paired summaries and direction counts. For each endpoint we report the paired delta, its median, and explicit win, tie, and loss counts, since a tie carries different information from a small delta of either sign; counts are stratified by task before pooling, which is why results like 5 wins, 9 ties, and 1 loss are reported as counts beside their means.

Scale-normalised error. Raw MSE is not comparable across targets of different magnitude, so paired private ratios (Eq. (47)) summarise treatment against control within a task, with non-inferiority gates stated on them, and the harness NMSE of Eq. (38) is the scale of the matched-trajectory and two-anomalies private gates, both at a relative margin of 0.05; Eq. (25) is the generator-level form the frozen calibration bands constrain.

Cross-seed uncertainty. Three seeds per task instance support per-seed reporting and direction counts, and no parametric confidence interval is fitted to three paired observations; where a block is small, the per-cell values are printed in full. Gates are stated as counts over pairs and evaluated only after the expected-cell check passes, and development runs that failed a frozen gate never enter confirmatory estimates.

D.4 LEAKAGE, CORRECTNESS, AND REGRESSION AUDITS

Three failure modes threaten the comparisons above: information that reaches a decision before it is legitimately available, a scoring or parsing defect that silently changes a reported number, and a run that is analysed while incomplete or after its code has drifted. The audits in this subsection are organised by those modes, run before or during every confirmatory launch, and abort the run or refuse the analysis when they fail. Their per-package instances are listed in Appendix E.2, and the frozen digests and manifests they check are recorded in Appendix F.2.

Parser and compiler regression tests. The DSL surface layer carries deterministic tests that must pass before any confirmatory cell executes, because a tokenizer defect changes fitted coefficients and every downstream metric. The suite checks grammar round-trips, infix nonlinear parsing and compilation, complete scientific-notation literals, and signed multiline scientific-notation evaluation; the capability-panel extension adds pre-extension validity, piecewise multi-equation execution, and recomputation of measured intermediates. The scientific-notation repair is part of the frozen post-fix pipeline, and results produced before it are excluded from the post-fix archive.

Metadata and role-label isolation. Ground-truth roles exist only in the offline benchmark and the scoring code. SYMBOLICREFIT receives no hidden-role labels, ORACLEQUERY returns measured values with no correctness or relevance verdict, and the public serialisation of a harness cell is scanned for offline vocabulary before the cell runs. In the two-anomalies package the structural cell audit lowercases the serialised public task, candidate bundle, initial evidence, initial state, and action menu, and fails the cell when any of eleven banned tokens appears, among them `ground_truth`, `correct_candidate`, `task_type`, `n_type`, `v_type`, `spurious`, and `accommodation`. The same audit checks that the auxiliary representation transition string appears exactly once, that the menu holds exactly one out-of-regime probe and at least two repeatable actions at equal cost, that the probe separates the correct candidate from every wrong one, and that all in-regime candidate errors satisfy the adequacy bound $\mathrm{NMSE} \leq 0.15$.

Semantic matcher boundary. Canonical resolution can leak ground truth if the matcher sees more than the discovery-visible inventory. RESOLVECOMPONENT therefore attempts exact identifier match, then lexical token overlap, then the alias table, and only then the LLM matcher, whose prompt is restricted to the question, the target, the visible inventory with aliases and descriptions, and evidence snippets. The visible inventory is set explicitly from the currently available columns before each resolution round, simulator role labels and graph metadata are excluded, and a match is accepted only above the configured confidence threshold of 0.72; unresolved proposals are recorded as unmatched with that reason.

Private-test and panel-identity checks. The sealed panel refuses release without a frozen fingerprint and refuses a second release under a different one; deterministic tests cover this behaviour, null private values on no-commit trajectories, private evaluation on observed panels only, and capability-panel release only after the final model freeze. Panel identity is checked at two levels. Within a pair, the analyzer requires exact equality of the initial state, candidate bundle, action menu, outcome schedule, and calibration hashes, and the matched-input gate demands a matched fraction of 1.0; the merged two-anomalies run records 1.0. Across reused panels, the accepted archive verifies exact row identifiers, coverage, provenance, and split-boundary compliance, and the streaming arms rebuild each panel from the same outer-step index and task offset used during discovery.

Manifest, resume, and source-digest checks. At launch the runner writes a manifest holding the frozen design, the model profile, the frozen cell list, the code revision the run started from, and a digest map over the eleven behavioural source modules (Eq. (55)). Every resume re-verifies three invariants against that manifest, the recomputed cell matrix, the profile, and the recomputed digest map, and aborts on any mismatch, so a run cannot continue silently under edited code or an edited

design; each resume is appended to the run's resume history. Per-cell completion is recorded together with the digest of the cell's own record, and a cell counts as complete only when the recomputed digest matches, so a truncated or edited cell is re-run instead of being pooled. When a run is split across two machines, the merge requires the two digest maps to be identical and refuses any cell whose record changed after completion.

Expected-cell-count and completeness gates. Gates are evaluated only on a complete matrix. The two-anomalies design freezes 30 expected cells, 24 in the main matrix and 6 in the forced-probe component, and the validity gate checks completeness, the calibration hash, the per-cell design audits, the matched-input fraction, the per-arm fallback cap of 0.25, and digest consistency across run directories before any endpoint is computed; the merged run reports 30 of 30 cells with per-arm fallback rates 0.0208 for epistemic control and 0 for the fit-guided and forced-probe arms. The matched-menu and diagnosis-stress analyses require all 24 cells, and the accepted archives verify 60 cells in each of the streaming-versus-static comparison, the layered-reference comparison, and the estimator-class diagnostic, and all 240 cells of the cumulative-evidence comparison. Before a launch the package unit tests run, followed by a dry run that re-derives the frozen calibration hash, the per-cell audits, the menus, and the full cell matrix; GPU availability is deliberately left unprobed so that the design check stays machine-independent.

Post-completion analysis-layer changes. Audit repairs applied after a run completes are recorded and scoped. In the two-anomalies package the analysis layer was changed after the confirmatory run to refuse with a named list of incomplete cells and to tolerate brief cross-machine visibility delay of completion markers, after a race appeared at merge time; no run record was altered. The erratum in the ordered-program study likewise changes only the completeness check for retained invalid-request evidence and alters no run cell, model output, threshold, private-test value, or gate. The cumulative-evidence archive records a deterministic analysis-only reconstruction of missing derived query-efficiency values from the canonical query-count definition, with raw outputs and decisions unchanged.

E Experiment details and supplementary results

This appendix holds the evidence record behind every number in Section 4, one subsection per claim area, including the failed dominance gate in full and the development lineage that preceded each confirmatory run.

E.1 Complete evidence inventory

Tables 11 and 12 record every completed experiment used as evidence in the paper. Main denotes direct use in Section 4; supporting denotes context for a main result; diagnostic denotes a valid experiment that does not sustain the intended capability claim; and boundary denotes evidence whose scope is deliberately narrower than the paper's central claim. Development runs that failed a frozen gate are retained in the artifact lineage but are not pooled with confirmatory estimates.

The released record labels each panel, tier, and study with a short artifact identifier, and those identifiers appear nowhere else in this appendix: the reference panel (`hard-v2`, P0), the capability panel (P1), the matched-trajectory suite (P2), the two-anomalies suite (P3), the split-integrity check (B03), the streaming-versus-static comparison (C01), the layered-reference comparison (C02), the estimator-class diagnostic (C03), the literature-exposure suite (C04), the cumulative-evidence comparison (C05), the numeric-preview ablation (C06), the basis-expansion study (C07), the multiple-candidate search study (C08), the joint-commitment study (C09), the deletion-safety audit (C10), the ordered-program structural study (C11), the proxy-contamination track (C12), the real-data feature-selection analogue (C13), the evidence-gated deletion comparison (P1.5), the matched-evidence comparison (A1-v3), the matched-menu comparison (P2-B), the reopening comparison (A2-v2), the diagnosis-stress comparison (C1-v2), and the component-revision comparison (P2-C2).

The two-anomalies suite has three frozen runs; Appendix E.6 records each run's prespecified revision and gate record, and Table 26 summarises the lineage.

Table 11: Complete evidence inventory for the prototype tiers and the two-anomalies suite. Results are aggregate summaries; the per-cell outputs and acceptance records are supplied with the reproducibility material.

Study	Scale	Paper-facing result	Use
Split-integrity check	1 task, 1 split	Sealed MSE $23789.829 \rightarrow 0.253$; hidden precision/recall 1.0; exact DSL match 0	Supporting protocol smoke test
Streaming-versus-static comparison	60 cells	Streaming and static mean private MSE differ by 0.0023; hidden F1 $0.969 \rightarrow 0.893$ and query efficiency $0.950 \rightarrow 0.713$	Supporting observational-robustness result
Layered-reference comparison	60 cells	Discovery, GT-parent refit, and GT-mechanism MSE are 35.254, 30.460, and 4.076	Main layered-recovery result
Estimator-class diagnostic	60 cells	Same-variable nonlinear probe gives no general advantage; GT-variable and GT-mechanism references remain distinct	Diagnostic only
Literature-exposure suite	108 cells	Alternating irrelevant exposure lowers hidden F1 $0.983 \rightarrow 0.822$; obvious distractors are null; median paired MSE deltas are zero	Supporting, confounded policy evidence
Cumulative-evidence comparison	240 cells	Cumulative vs. delta-only hidden F1 is 0.928 vs. 0.834; over-discovery is 0.300 vs. 0.650	Supporting evidence-state result
Numeric-preview ablation	8 cells	Numeric preview changes neither variable recovery nor private MSE on four ceiling-level tasks	Appendix null result
Basis-expansion study	30 cells	Augmented terms: 5 wins, 9 ties, 1 loss; external interaction selected in 3/3 materials seeds, MSE ratio 0.018	Main task-specific evidence for principle P3
Plural-candidate search study	6 cells	Non-first candidate selected and private MSE improved in 3/3; aggregate multi/single ratio 0.594	Main evidence for principle P3
Joint-commitment study	12 cells	Exact joint recovery 4/6 vs. 6/6; unresolved no-commit exposures 18 vs. 0	Main evidence for principle P3
Deletion-safety audit	247 exclusions + 24 cells	LOVO marks 15 GT parents removable; predictive-only/evidence-gated GT-variable losses are 23/0	Main negative safety evidence
Ordered-program structural study	12 cells	Explicit treatment reaches full-edge F1 1, SHD 0, and intermediate recall 1 in 6/6; private-MSE ratio 0.943	Main structural evidence for principles P1 and P3
Proxy-contamination track	60 + 12 cells	Initial proxy exposure raises retention 4/30 to 20/30; declared typed replacement removes the proxy in 6/6	Boundary track
Real-data feature-selection analogue	3 splits, 4 conditions	Progressive nine-feature MSE 0.0634 vs. initial 0.0784 and random 0.0740; selected-set Jaccard 0.343	Boundary real-data analogue
Two-anomalies comparison	30 cells $\times$ 3 frozen runs	Final run: baseline false retention 3/6 vs. 0/6; all control-arm gates pass; dominance factors 27–85 where the baseline is fooled; V-dominance gate 3/6	Main retention-direction evidence; one secondary gate failed and reported

Table 12: Frozen confirmatory evidence of the matched-trajectory suite. All five blocks pass their complete prespecified gate sets; 108 cells form 54 matched pairs.

Block	Pairs	Fit-guided policy	Epistemic control	Gate
Matched evidence	12	Typed response 0; defeated-account retirement 0	Typed response 1; retirement 1; both arms final-correct 1	13/13
Matched menu	12	First discriminating action 0.75; cost 1.25; support-seeking 0.125	First discriminating action 1; cost 1; support-seeking 0	11/11
Reopening	6	Reopening, withdrawal, rollback, and final correctness 0 after decisive evidence	All four outcomes 1; both arms stable before decisive evidence	11/11
Diagnosis stress	12	Response accuracy 0.5; unnecessary repair 0.5	Response accuracy 1; unnecessary repair 0; no-commit 3/3; replication 3/3	13/13
Component revision	12	Correct proposal and final account 0.5; mean private NMSE 0.0187	Correct proposal and final account 1; useful removal, retention, coherence 1; NMSE 0.0052	13/13

E.2 INTEGRITY AND REPRESENTATION-FAILURE LOCALISATION

The frozen results in Tables 11 and 12 rest on an integrity layer that constrains what the discovery process may read at each decision and binds every released number to one audited artifact. This subsection documents that layer, then reports the diagnosis-stress comparison, in which four induced failure families test whether the response matches the diagnosed failure; its failed first formulation is retained as labelled development evidence, and its second alone supplies the confirmatory estimate of Section 4.2.

Prohibited-access tests. The information-boundary audits of Appendix D.4 run before any confirmatory cell executes, and the controller prompts of the harness packages carry no expected gap, no correct-candidate identity, and no private outcome. Candidate order varies so list position cannot encode the answer, and a fit-guided action that contradicts the offline warranted action still validates, so the validator holds no answer key.

Private-test release fingerprints. In the matched-menu comparison the 512 private rows per task are physically absent during discovery and released only against a fingerprint of the frozen final state,

$$\phi = \mathrm{sha256}\big(\text{arm},\ \text{task},\ \text{seed},\ H_{\text{cand}},\ H_{\text{menu}},\ H_{\text{out}},\ \langle (a_r, \mathcal{S}_r)\rangle_{r \le R},\ \mathcal{S}_{\text{final}}\big), \tag{43}$$

which hashes the frozen bundles together with every selected action and reduced state, so each released private value is bound to exactly one audited trajectory. The capability-panel cells write a machine-checked private-test audit flag, and all 60 cells behind the four capability studies pass it.

Source hashes and expected-cell checks. Each frozen package ships a hash manifest over its designs, modules, and records; one recorded erratum in the ordered-program study changes only a completeness check and no run cell, output, threshold, or gate. Expected-cell checks run before any gate is evaluated, matched-input gates require exact hash equality of initial state, candidate bundle, and evidence trace within every pair, and the observed matched fraction is 1.000 in all five matched-trajectory blocks.

Four induced representation-failure families. The diagnosis-stress comparison crosses two warranted-repair families with two no-repair controls (Table 13); both arms receive the same deterministic appraisal, which carries candidate separation, the proposed transition type, and measurement reliability, and no task family, answer key, or private outcome.

Table 13: The four diagnostic families of the diagnosis-stress comparison. The visible-appraisal column lists the generic typed triple (candidate separation; proposed transition type; measurement reliability) that both policy arms observe; the warranted response is attached only in offline evaluation after the action is frozen.

Task	Induced limitation	Visible appraisal	Warranted response
Alloy parent diagnosis	Missing measured parent (missing-variable family): a phase-state coordinate outside $\mathcal{O}$ carries part of the mechanism	decisive; measured-quantity transition; reliability not flagged	Accept the variable-level repair
Thermal functional form	Inadequate functional form (functional-family family): an additive account cannot express a regime-dependent slope	decisive; functional-language transition; reliability not flagged	Accept the mechanism-level repair
Equivalent accounts	Observationally equivalent live accounts (observational-equivalence family): a ratio account and a saturating account agree inside the sampled regime	practically indistinguishable; no transition; reliability not flagged	No-commit; keep both accounts live
Sensor replication	Unstable measurement process (measurement-unreliability family): technical replicates disagree and the sensor offset shifts across calibration batches	decisive; no transition; reliability unsettled	Request clean replication before any representation change

The first formulation as development evidence. The first formulation completed all 24 cells and failed 4 of its 14 checks because one contract field conflated the initial limitation with the residual state after the response, rejecting otherwise correct repairs; the accepted formulation splits that field, and every surviving threshold was tightened while no outcome threshold was relaxed (Table 31).

Confirmatory diagnosis-stress outcomes. The frozen matrix crosses the four instances with fresh seeds and two matched policies at temperature 0. Both arms name the initial limitation correctly in 12

of 12 cells. The fit-guided policy performs the six warranted repairs and also changes the account in all six no-repair controls, for response accuracy 0.500 and unnecessary-repair rate 0.500; epistemic control keeps all six repairs, makes 3 of 3 no-commit decisions under observational equivalence, and issues 3 of 3 replication requests under measurement unreliability, for accuracy 1.000 and unnecessary repair 0.000. Claim-ceiling compliance is 1.000 in both arms and the fallback rate is one cell of twelve in each, under the 0.25 cap. The result excludes an always-repair shortcut within these four pre-specified families under matched inputs, and it supports no claim of unrestricted diagnosis.

Acceptance checks across the five blocks. Table 14 compresses the five prespecified gate sets to their outcomes; the full per-gate tables, with every threshold and observed value, ship with the release. Table 15 expands the two-anomalies run in full, because it contains the one failed gate.

Table 14: Prespecified acceptance checks of the five matched-trajectory comparisons. Every block requires complete cells, hash-matched pairs at fraction 1.0, treatment minima of 0.75 to 1.0 on its control endpoints, caps of 0 to 0.05 on unsupported claims, a 0.25 controller-fallback cap, and, where defined, private-test non-inferiority at relative margin 0.05. All gates pass in all five blocks.

Comparison	Gates	Observed highlights
Matched evidence	13/13	typed repair and retirement 1.000 against 0.000; retirement improvement $+1.000$; private non-inferiority 12/12
Matched menu	11/11	first discriminating action 1.000 against 0.750; mean actions to decisive evidence 1.00 against 1.25; identical private NMSE in 12/12 pairs
Reopening	11/11	reopening, withdrawal, rollback, and final correctness 1.000 against 0.000; stability before decisive evidence 1.000 in both arms
Diagnosis stress	13/13	response accuracy 1.000 against 0.500; unnecessary repair 0.000 against 0.500; ceiling compliance 1.000
Component revision	13/13	useful removal, warranted retention, and coherent revision 1.000; mean private NMSE 0.0052 against 0.0187

Table 15: The twelve prespecified acceptance checks of the confirmatory two-anomalies run (the third frozen run; 30 cells, comprising 24 main-matrix cells that form 12 matched pairs and a 6-cell forced-probe component), grouped into the four gate sections of the merged report, with their frozen thresholds and observed values. Eleven checks pass; the sealed V-dominance check returns 3 of 6 pairs against a frozen minimum of 5, so the private-test section and the run as a whole are recorded as FAIL on that one secondary gate.

Section	Check	Threshold	Observed
Validity	Expected cells complete	$= 30$	30
Validity	Source digests recorded by the two execution hosts agree	exact equality	identical
Validity	Matched within-pair hashes of the initial state, candidate bundle, action menu, outcome schedule, and calibration artifact	fraction $= 1.0$	1.000
Validity	Frozen calibration hash carried by every cell	exact equality	30/30
Validity	Per-cell design audit	all PASS	30/30
Validity	Controller fallback rate in the epistemic-control, fit-guided, and forced-probe arms	≤ 0.25 each	0.0208, 0, 0
Control arm	Auxiliary account rejected on V-type cells, excluding fallback cells	≥ 5 of 6	6/6
Control arm	Auxiliary account retained on every non-fallback N-type cell	all, with ≥ 5 such cells	5/5
Control arm	Bounded verification cost on N pairs: actions to first commit at most the fit-guided count plus 1	≥ 5 of 6 pairs	6/6
Comparative	V-type false-retention advantage over the fit-guided policy	delta ≥ 0.50	$+0.500$
Private test	Sealed mixed-regime V dominance: control NMSE at most $0.95\times$ the fit-guided NMSE	≥ 5 of 6 pairs	3/6
Private test	Sealed mixed-regime N non-inferiority: control NMSE at most $1.05\times$ the fit-guided NMSE	6 of 6 pairs	6/6

E.3 Inquiry, robustness, and reference decompositions

This subsection supplies the complete condition-by-endpoint matrices behind the inquiry-related results of Section 4.2: the streaming-versus-static acquisition comparison, the literature-exposure suite, the cumulative-evidence comparison, and the matched-menu comparison with its action menu, costs, selected actions, and downstream responses. It then separates the selector-only estimate from the policy-package effects of the matched-trajectory blocks, and closes with the layered-reference decomposition, the estimator-class probe, and the numeric-preview ablation.

Static and streaming observational acquisition. All ten reference-panel tasks run under seeds 0 to 2 in two acquisition regimes, 60 runs and 30 pairs, sharing profile, literature, budgets, and commit logic; static keeps one panel, streaming redraws one before each exposure. Private prediction is exactly balanced at 15 pairs each way, while hidden-variable F1 drops by 0.076, query efficiency by 0.237, and perfect hidden recovery from 26 to 16 of 30. The balanced win count is descriptive rather than an equivalence test. Structural recovery and query efficiency decrease under this observational acquisition schedule; no interventional claim follows.

The literature-exposure suite. Nine literature policies cross four reference-panel tasks and seeds 0 to 2, giving 108 sessions with the pipeline fixed and the exposure schedule as the only intervention. With H the hidden parents, D_{final} the acquired identifiers retained in the final account, and Q the requested identifiers,

$$P_{\text{hid}} = \frac{|D_{\text{final}} \cap H|}{|D_{\text{final}}|}, \qquad R_{\text{hid}} = \frac{|D_{\text{final}} \cap H|}{|H|}, \qquad F1_{\text{hid}} = \frac{2\,P_{\text{hid}} R_{\text{hid}}}{P_{\text{hid}} + R_{\text{hid}}}. \tag{44}$$

The nine policies range from a clean progressively informative reference, through redundant, rival, cross-domain, and shared-vocabulary mixtures, to two ground-truth curricula and an alternating-irrelevant schedule; Table 16 gives the final recovery under each.

Table 16: Final hidden-variable recovery per literature policy over 12 task-seed cells (four tasks, three seeds). Precision, recall, and F1 follow Eq. (44); the SD is the sample standard deviation over cells; the paired $\Delta F1$ is taken against the clean ground-truth policy within the same task and seed.

Policy	Precision	Recall	F1 $\pm$ SD	Perfect cells	Paired $\Delta F1$
Clean ground truth	0.972	1.000	0.983 ± 0.058	11/12	
Original hard mixed	1.000	1.000	1.000 ± 0.000	12/12	$+0.017$
Irrelevant cross-domain	1.000	1.000	1.000 ± 0.000	12/12	$+0.017$
Redundant and seen	1.000	1.000	1.000 ± 0.000	12/12	$+0.017$
Rival mechanisms	0.972	1.000	0.983 ± 0.058	11/12	0.000
Shared vocabulary, non-SCM	0.972	1.000	0.983 ± 0.058	11/12	0.000
Curriculum, large drop	0.792	1.000	0.872 ± 0.119	5/12	-0.111
Curriculum, single drop	0.764	1.000	0.856 ± 0.113	4/12	-0.128
Alternating irrelevant	0.736	0.958	0.822 ± 0.174	4/12	-0.161

Alternating irrelevant exposure is the only policy with an aggregate recall loss, and it carries the highest off-target request fraction, 0.319, and the largest post-recovery over-acquisition; both ground-truth curricula keep recall at 1.000 while precision falls to 0.792 and 0.764, so their dominant error is retaining provisional proxies after the mechanism is complete. The median paired private-MSE difference is zero for every policy, so prediction exposes none of these structural errors. Document and token budgets are unmatched across policies and the alternating schedule changes relevance and continuity together, the declared confound of Section 4.5.

Cumulative and delta-only evidence states. The accepted matrix crosses the ten tasks, three seeds, four literature policies, and two exposure modes, 240 cells and 120 pairs; within a pair everything is shared except whether decisions see the cumulative typed evidence or only the newly exposed batch. Cumulative context raises hidden F1 from 0.834 to 0.928, halves the irrelevant-proposal rate, and halves over-discovery after full recovery from 0.650 to 0.300, winning 50 pairs, losing 9, and tying 61. The advantage is schedule-dependent, largest under adversarial exposure and slightly reversed under the self-contained single-drop curriculum, and the design changes retained context and prompt length together, so the gain cannot be attributed to factual continuity alone.

The matched-menu comparison. Four controlled tasks cross two selector arms over seeds 401, 503, and 607, 24 cells and 12 pairs; each trajectory starts from a frozen panel and two executable candidates whose initial evidence is insufficient at the registered ambiguity margin. Both arms receive the same three equal-cost actions, a discriminating measurement or regime, a same-scope replication, and a precision audit, in counterbalanced order with neutral identifiers; at most two actions may be taken, a shared deterministic appraisal commits only on a decisive separation, and with $\text{role}(a_r)$ the sealed offline role of the r-th selected action,

$$N_{\text{dis}} = \min\{r : \text{role}(a_r) = \text{discriminating}\}, \qquad c_{\text{dis}} = \sum_{r=1}^{N_{\text{dis}}} c(a_r). \tag{45}$$

The only textual difference between the arms is one selection-policy paragraph, relevance and support against expected outcome difference across the live candidates (Appendix F.1).

Table 17: The action menu of the matched-menu comparison. All three actions are feasible, carry equal cost, and appear in a seed-counterbalanced order; the offline role labels never appear in the public menu and are used only in sealed evaluation. The discriminating action is instantiated per task as a targeted measurement of the proposed coordinate (alloy), a targeted measurement of the proposed transport intermediate (reactor), or a wider temperature or resistance regime (thermal, transport).

Public action description	Semantic type	Offline role	Cost
A targeted measurement or new observational regime whose candidate-specific feasible outcomes differ	`targeted_measurement` or `new_observational_regime`	discriminating	1.0
A same-scope independent replication that can support the provisional account without distinguishing the rivals	`same_scope_replication`	support	1.0
A higher-precision audit of the already measured quantities in the current operating scope	`measurement_precision_audit`	generic	1.0

Selected actions and downstream responses. Table 18 reports the full matrix. The fit-guided selector takes a non-discriminating action first in 3 of 12 cells, each time the same-scope replication, and reaches the discriminating action on its second attempt; the uncertainty-resolving selector chooses it first in 12 of 12, so mean actions to decisive evidence fall from 1.250 to 1.000. In the main-model matched-menu comparison, both arms obtain decisive evidence within budget in every run, reach the correct model, and pass the claim-ceiling check; private-test NMSE is identical within all 12 pairs.

Table 18: Core endpoints of the matched-menu comparison over 24 cells; the full selection and response matrix ships with the release.

Endpoint	Fit-guided	Epistemic control
First action discriminating (12 cells)	0.750	1.000
Mean actions to decisive evidence	1.250	1.000
Support-seeking action rate	0.125	0.000
Final account correct	1.000	1.000
Ceiling compliance; unsupported causal claims	1.000; 0	1.000; 0
Mean private-test NMSE	0.006993	0.006993

Selector-only estimate and policy-package effects. The arms differ in one selector paragraph while menu, costs, outcomes, budget, appraisal, and release rule are shared, so the first-action gain and the cost reduction are attributable to inquiry selection alone, and the identical final accounts and private predictions show that nothing downstream moved. Effects that need a different response policy are estimated by the other four comparisons, whose fixed evidence sequences in turn measure no inquiry efficiency, so neither estimate can inflate the other.

Layered reference decomposition. The layered-reference comparison replaces one scientific input at a time in the accepted artifacts of the streaming comparison, rerunning no discovery: E_D is the frozen discovered account, E_V the same estimator supplied with the exact parents, and E_M the executed generating equation, giving the gaps

$$G_{\text{disc}} = E_D - E_V, \qquad G_{\text{func}} = E_V - E_M. \tag{46}$$

Table 19 gives the 60-cell matrix. Both gaps are positive in aggregate and G_{func} dominates, 26.38 of the 31.18 aggregate distance; the paired counts qualify the means: the generating mechanism beats the parent refit in 47 of 60 cells, and the per-stage medians, 1.76, 1.54, and 1.37, sit far below the outlier-driven means, the descriptive median summary reported in Section 4.4. Differences of marginal medians are not paired error decompositions. The parent arm is an oracle-input diagnostic that inherits the estimator's search class, budget, and finite-sample behaviour. Executing the generating mechanism additionally supplies its functional form and parameters, so the residual gap cannot be attributed solely to failed mechanism search.

The random-feature diagnostic and its status. The estimator-class diagnostic crosses the input set with a frozen random-feature `tanh` ridge probe over the same 60 cells. The symbolic account has lower private error than the probe on the same variables in 53 of 60 cells; replacing discovered

Table 19: Layered reference matrix over 60 cells. E_D is the frozen discovered account, E_V the ground-truth-parent symbolic refit, and E_M the executed generating mechanism, all on the sealed private test; gaps follow Eq. (46). Target scales differ across tasks, so the paired counts in the text accompany these arithmetic means.

Regime	Cells	Mean E_D	Mean E_V	Mean E_M	Mean G_{disc}	Mean G_{func}
Static fixed panel	30	35.2530	34.0747	3.4656	1.1783	30.6092
Streaming observational	30	35.2553	26.8457	4.6861	8.4096	22.1596
Overall	60	35.2541	30.4602	4.0758	4.7939	26.3844

variables with the true parents under the same probe lowers the mean from 35.85 to 15.72, and executing the generating mechanism lowers it further, positive in all 60 cells. It is reported as a probe and supports no claim about model classes; the residual error is jointly attributable to finite samples, capacity, regularisation, and misspecification.

Numeric preview inside the hypothesis prompt. The numeric-preview ablation removes one numeric block from the hypothesis-refinement prompt and changes nothing else, over four reference-panel scenarios at one seed. Every paired difference is exactly zero, in hidden-variable F1 and in private-test MSE, while the prompt logs confirm the manipulation was active in every run. All four tasks sit at the recovery ceiling with F1 1.000 in both arms, so the null bounds these tasks and leaves smaller intermediate differences invisible; the archived run also predates one leakage repair shared identically by both arms, which leaves the paired contrast intact and the absolute figures unsuitable as a standalone benchmark result.

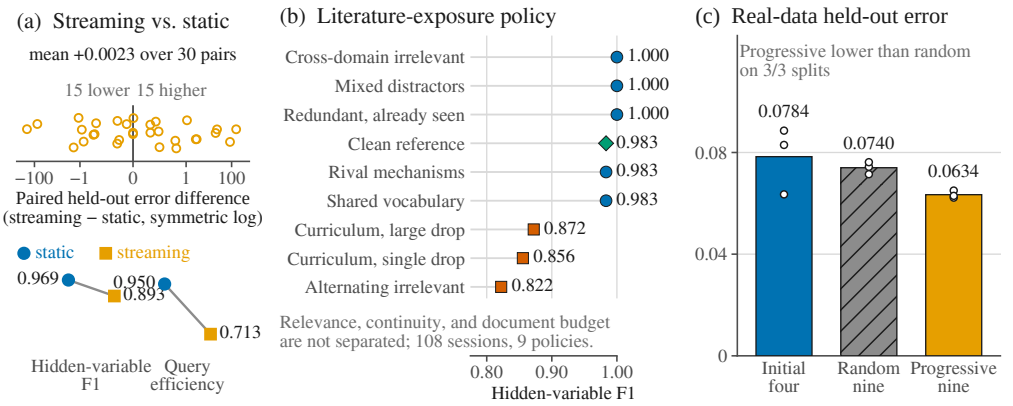


Figure 6: Robustness checks and boundary evidence. (a) Streaming acquisition against static exposure: the paired held-out error differences over 30 pairs centre on zero, with 15 lower and 15 higher, while hidden-variable F1 falls from 0.969 to 0.893 and query efficiency from 0.950 to 0.713. (b) Hidden-variable recovery by literature-exposure policy over 108 sessions; the alternating-irrelevant policy is the worst at 0.822 against 0.983 for the clean reference, and the study does not separate relevance, continuity, and document budget. (c) The real-data analogue: held-out test MSE for the progressive nine-feature set against the initial set and a size-matched random subset, with the three splits shown as points.

E.4 Multiple-hypothesis search

This subsection gives the per-cell records behind Section 4.4: each term outside the predefined basis with its provenance, the candidate-set diversity of the multiple-candidate search, the component exposures under joint commitment, and the complete edge results of the ordered-program structural study with their confirmatory lineage. Throughout, the paired private ratio and the aggregate ratio are

$$\rho = \frac{E^{\text{treat}}_{\text{priv}}}{E^{\text{ctrl}}_{\text{priv}}}, \qquad \rho_{\text{agg}} = \frac{\sum_{\text{cells}} E^{\text{treat}}_{\text{priv}}}{\sum_{\text{cells}} E^{\text{ctrl}}_{\text{priv}}}, \tag{47}$$

with E_{priv} the sealed private-test error of the frozen final model and values below one favouring the treatment arm.

Proposed terms outside the predefined basis. Five capability tasks, seeds 0 to 2, and two candidate-source policies give 30 cells and 15 pairs; the fixed-basis arm draws from the predefined library alone, the hybrid arm adds LLM-proposed executable terms, and everything else is matched. A selected term counts as outside-library only when it is absent from the fixed library and passes parsing, compilation, and finite-output checks; LLM provenance alone is insufficient, because the generator can re-derive terms the grammar already contains. Table 20 gives the paired summary: the one reproducible external discovery is the materials product interaction, selected in all three seeds with fitted coefficients between 83.29 and 85.66, against fixed-basis models stuck near private MSE 900, a ratio of 0.018. The crop task improves at mean ratio 0.478 while selecting no external term, so that gain is not open-ended invention, and the remaining tasks tie. The study exercises term generation on a fixed acquisition trace. The added terms are already executable in the DSL; it contains neither autonomous diagnosis of language inadequacy nor a language change.

Table 20: Task-level paired results of the basis-expansion study over 15 pairs (five tasks, seeds 0 to 2). The ratio is the mean of the per-pair values of Eq. (47) with the LLM-augmented hybrid as treatment; wins, ties, and losses count pairs in which the hybrid arm has lower, equal, or higher sealed private MSE; the last column counts cells that select a term outside the predefined basis.

Task	Pairs	Mean ratio	Wins/ties/losses	External selected
Materials interaction strength	3	0.018	3/0/0	3/3
Crop piecewise stress	3	0.478	2/1/0	0/3
Building threshold energy	3	1.001	0/2/1	0/3
Reaction nested ratio	3	1.000	0/3/0	0/3
River joint conveyance	3	1.000	0/3/0	0/3

Candidate-set diversity in plural search. On the battery competing-DSL task, first-candidate identity is verified in 3 of 3 pairs, the plural arm generates 12.33 unique executable candidates per cell against 5.00, selects a non-first candidate in every cell, and records eight validation-selected rescues, each surviving the sealed test (Table 21); private MSE falls in all three pairs with aggregate ratio 0.594 and variable F1 at 1.0 in both arms, so the whole gain is mechanism selection. The scope is one purpose-built task and three seeds.

Table 21: Candidate-set diversity and paired outcomes of the multiple-candidate search study on the battery competing-DSL task. Unique candidates count distinct executable DSLs generated in a cell; distinct steps, non-first selections, and rescues count inner-loop steps; MSE columns are sealed private-test errors, and the mean-row ratio is ρ_{agg} of Eq. (47).

Seed	Unique (single)	Unique (plural)	Distinct steps	Non-first selections	Rescues	Single MSE	Plural MSE	Ratio
0	5	13	5	4	3	1.5639	1.0009	0.640
1	5	12	4	4	4	2.1976	1.4600	0.664
2	5	12	4	3	1	2.8398	1.4629	0.515
Mean	5.00	12.33				2.2004	1.3079	0.594

Component exposures under joint commitment. On the two joint-only tasks, where the necessity audit shows each singleton moves validation error by under 0.02 per cent while the pair improves it by about 98 per cent, the joint-component arm makes exactly one exact joint commit per cell with zero no-commit exposures, while the singleton arm accumulates 18 no-commit exposures and fails both seed-2 cells with an empty acquired set. Private MSE is identical where both arms recover the pair and improves exactly where the singleton arm fails. The study tests joint proposal and acquisition under the configured improvement rule. The singleton arm succeeds in four of six runs, so the numerical singleton-improvement audit is not an impossibility result for the full sequential procedure. Unrestricted group selection is untested.

Complete edge results of the ordered-program structural study. The confirmatory matrix crosses the two multi-hop tasks with fresh seeds 12 to 14 and two representations, scored on the directed

edges of the full benchmark graph by

$$P_E = \frac{|\widehat{E} \cap E^\star|}{|\widehat{E}|}, \qquad R_E = \frac{|\widehat{E} \cap E^\star|}{|E^\star|}, \qquad F1_E = \frac{2\,P_E R_E}{P_E + R_E}, \tag{48}$$

$$\mathrm{SHD} = |E^\star \setminus \widehat{E}| + |\widehat{E} \setminus E^\star|, \tag{49}$$

with reversed edges counted separately; the benchmark graphs hold nine and six directed edges. Table 22 gives the per-cell record: every explicit-program cell reproduces its benchmark graph exactly, while the flat arm keeps the true target edges and introduces direct target shortcuts while omitting local equations for intermediates. The reported intermediate-edge score has a definition discrepancy noted beside Eq. (37); the full-edge counts are retained as reported. On the sealed test the explicit program is lower in five of six pairs with aggregate ratio 0.943 against the 1.10 boundary, and all 11 prespecified gates hold.

Table 22: Complete per-cell edge results for the confirmatory run of the ordered-program structural study (seeds 12 to 14). Pred., Corr., Miss., Extra, and Rev. count predicted, correct, missing, extra, and reversed directed edges against the benchmark graph, which holds nine edges for air quality and six for protein expression; F1 and SHD follow Eqs. (48) and (49); Int. rec. is the recorded intermediate-edge score (see the definition discrepancy beside Eq. (37)); MSE is private-test error.

Task	Seed	Arm	Pred.	Corr.	Miss.	Extra	Rev.	F1	SHD	Int. rec.	MSE
Air quality PM2.5	12	Flat	8	3	6	5	0	0.353	11	0.250	2.4349
Air quality PM2.5	12	Explicit	9	9	0	0	0	1.000	0	1.000	2.3113
Air quality PM2.5	13	Flat	9	3	6	6	0	0.333	12	0.250	2.6331
Air quality PM2.5	13	Explicit	9	9	0	0	0	1.000	0	1.000	2.4897
Air quality PM2.5	14	Flat	8	3	6	5	0	0.353	11	0.250	2.3188
Air quality PM2.5	14	Explicit	9	9	0	0	0	1.000	0	1.000	2.2802
Protein expr.	12	Flat	4	3	3	1	0	0.600	4	0.400	0.5348
Protein expr.	12	Explicit	6	6	0	0	0	1.000	0	1.000	0.4019
Protein expr.	13	Flat	4	3	3	1	0	0.600	4	0.400	0.4781
Protein expr.	13	Explicit	6	6	0	0	0	1.000	0	1.000	0.4817
Protein expr.	14	Flat	5	3	3	2	0	0.545	5	0.400	0.4030
Protein expr.	14	Explicit	6	6	0	0	0	1.000	0	1.000	0.3362

Confirmatory status and development lineage. Only the fifth iteration, on fresh seeds 12 to 14, is confirmatory evidence for this study; the first four are development or diagnostic lineage. Table 23 records each iteration with the gate that stopped it. The frozen evidence index rules out pooling the four earlier iterations with the confirmatory one or enlarging the final seed count with them; they are recorded to expose the development path and prevent retrospective hiding of failed gates. The claim itself is bounded to measured intermediates and a bounded program language, and it establishes no identification of arbitrary latent structure from observational data (Section 4.4).

Table 23: Iteration lineage of the ordered-program structural study. Each row is one complete two-task run on fresh seeds, numbered in order of execution; failed iterations are retained as labelled development or diagnostic evidence and are never pooled with the confirmatory cells.

Iteration	Fresh seeds	Diagnostic outcome	Evidence status
1	0–2	Initial confirmatory attempt exposed incomplete explicit retention and structure	Development only
2	3–5	Explicit-retention and structural gates failed	Diagnostic only
3	6–8	Structural gates passed; aggregate private-MSE ratio 1.213 failed non-inferiority	Diagnostic only
4	9–11	Structural gates passed; acquisition closure 0.929 and private-MSE ratio 3.598 failed	Diagnostic only
5	12–14	All 11 prespecified completion, integrity, structure, fallback, and private-test gates pass	Final confirmatory evidence

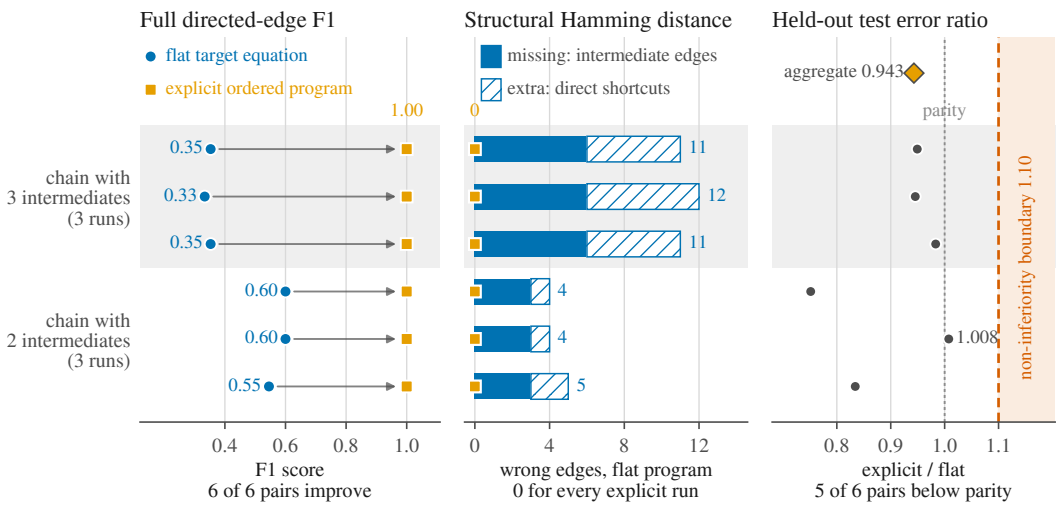


Figure 7: Structural recovery on the two multi-hop tasks, six fresh pairs over a chain with three measured intermediates and a chain with two. Left: full directed-edge F1 rises from the flat target equation to the explicit ordered program in all six pairs and reaches 1.00 in every one of its cells. Middle: the structural Hamming distance of the flat program splits into missing edges, associated with the omitted intermediate equations, and extra direct shortcuts into the target, while the explicit program scores 0 with intermediate-edge recall 1.00. Right: held-out test error ratio per pair, with the aggregate 0.943 against the prespecified tolerance of 1.10.

E.5 Falsification and revision

This subsection assembles the deletion evidence behind Section 4.3 and the positive revision evidence behind Section 4.2, and closes with the counted definition of a successful removal that all five frozen records share.

The leave-one-variable-out exclusion accounting. Stage A applies the frozen post-fit diagnostic of Eq. (42) to the 60 accepted streaming-comparison cells, 247 exclusion tests in all, counting variable instances inside cells. Of the 25 flagged instances, 15 are ground-truth parents, 7 of them initially hidden; the diagnostic catches only 10 of the 22 distractor instances, and prediction gives no warning at the point of deletion, since applying every flagged deletion moves mean private-test MSE by $+0.15$ overall and lowers it under streaming panels. Of the 15, 7 are initially hidden parents and 8 initially observed ones; Table 24 decomposes the counts by scenario, and crop failure, materials tensile strength, and river daily flow supply 14 of the 15 dangerous prunes.

Table 24: Exclusion tests of the deletion-safety audit by scenario over six cells each (three seeds $\times$ two acquisition regimes).

Scenario	Exclusion tests	Flagged removable	Dangerous prunes
Air quality PM2.5	20	0	0
Battery capacity loss	26	2	0
Building energy	25	1	0
Concrete strength	26	1	0
Crop failure	25	7	7
Materials tensile strength	28	4	4
Protein expression	32	5	1
Reaction yield	25	1	0
River daily flow	16	3	3
Turbine efficiency	24	1	0

The failed validation-only revision lineage. The validation-only revision run put a grouped controller with validation-only authorisation inside the full pipeline, 48 cells and 24 pairs over the capability tasks. It improved target-variable F1 and edge scores and matched private-test MSE, and it

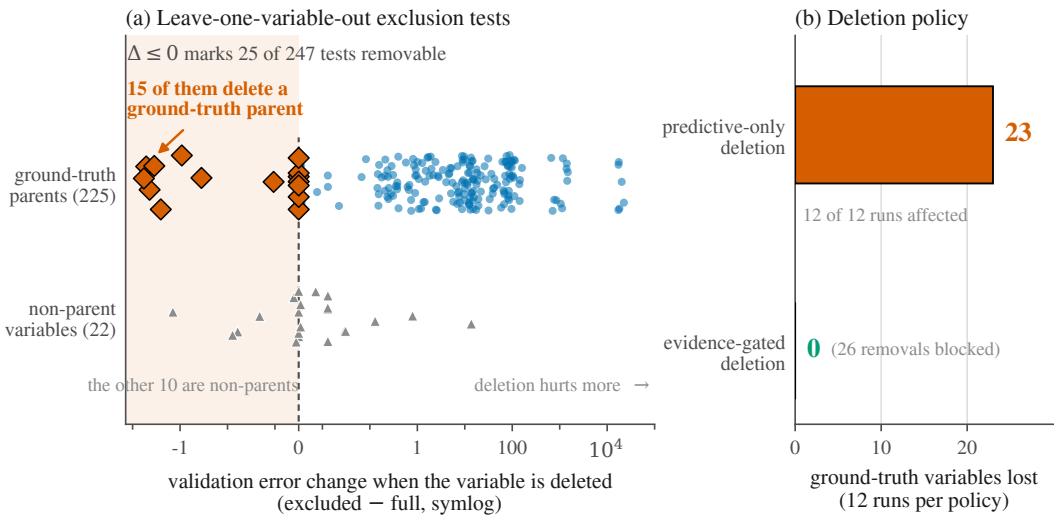


Figure 8: The deletion failure direction: predictive-only deletion loses 23 ground-truth variables, the evidence-gated policy none.

also cut hidden-variable F1 from 0.805 to 0.476, losing 34 ground-truth variable instances and one edge, so the zero-loss safety gates failed. The frozen root cause is exact: the pruning assessment had already labelled predictive redundancy diagnostic-only, and the session controller deleted such groups anyway, so diagnosis and authorisation were one step in effect; the run stays diagnostic lineage.

Evidence-gated conservative non-removal. The evidence-gated deletion comparison keeps the same proposals and changes only the authorisation rule: a removal group G is authorised only when every member carries explicit counterevidence eligibility,

$$\mathrm{Auth}(G) = \prod_{x \in G} \mathbf{1}\big[x \in \mathcal{C}_t\big], \tag{50}$$

and falsification is disabled by design, so $\mathcal{C}_t = \varnothing$ and every non-empty group must be blocked. Over 24 cells and 12 pairs the predictive-only arm accepts 22 deletions and loses 23 ground-truth instances; the gated arm blocks a proposal in all 12 cells and loses nothing, while mean private-test MSE separates the arms by only 0.054 (Table 25). All six prespecified checks pass. What this establishes is conservative non-removal under an empty eligibility set; useful pruning needs an eligibility set that counterevidence can populate, which the component-revision comparison supplies.

Table 25: Predictive-only against evidence-gated deletion authorisation over 24 cells and 12 pairs on four capability-panel tasks, with the same leave-one-variable-out and resampling proposals in both arms and falsification disabled.

Outcome over 12 pairs	Predictive-only	Evidence-gated
Accepted deletions	22	0
Ground-truth variable losses	23	0
Ground-truth edge losses	0	0
Cells with a blocked predictive proposal		12
Pairs with improved ground-truth retention		12
Mean private-test MSE	4.2738	4.3282

The matched-evidence comparison. Four controlled instances, labelled as two representation gaps and two hypothesis-language gaps in the harness, cross fresh seeds and two policies, 24 cells and 12 pairs. Each supplies two frozen executable accounts and a fixed three-event evidence sequence whose final event raises the claim ceiling. The labels describe the prescribed revision in the scenario; the agent selects among supplied accounts rather than inventing a new language. Within a pair the arms share the data, the counterbalanced bundle, the evidence trace, the appraisal, and one deterministic reducer, and the matching gate confirms identical hashes in 12 of 12 pairs. Both arms

end final-correct with identical private NMSE inside every pair; the difference is the scientific state. The fit-guided policy leaves the rejected list empty and the gap unresolved in every cell, so its typed repair and retirement are 0.000, while epistemic control reaches 1.000 on both in 12 of 12 pairs, and all thirteen gates pass. The estimand is the package-level effect of the declared policy, and the block attributes nothing to individual components, establishes no language invention, and measures no inquiry efficiency, since the evidence sequence is fixed.

Reopening traces after a released outcome. The reopening comparison starts each trajectory from an already released outcome and delivers a non-decisive monitoring event, then decisive counterevidence, 12 cells and 6 pairs. Both arms stay stable through the non-decisive event; after the decisive one, epistemic control reopens the loop, moves the defeated candidate to the rejected list, marks the released claim withdrawn, invalidates and recomputes the dependent fit summary, and advances the representation version, reaching 1.000 on every reopening endpoint, while the fit-guided arm ends append-and-retain in all six cells. Paired private NMSE ratios run from 0.0093 to 0.0170 and all eleven gates pass; the full per-round traces ship with the release.

Useful-removal and warranted-retention traces. The component-revision comparison crosses two removal tasks, whose challenged component fails to transfer to the later regime, and two retention tasks, whose component only looks redundant in the initial regime, with three seeds and two policies, 24 cells and 12 pairs. The fit-guided arm is append-only; the treatment arm may request a removal, and a deterministic discovery-visible gate authorises it only when the component-free account is non-inferior on every frozen resample,

$$\mathrm{Auth} = \mathbf{1}\big[a = \texttt{remove_component}\big] \cdot \Big(\textstyle\prod_{r=1}^{3} \mathbf{1}\big[\delta_r \le \tau\big]\Big) \cdot \mathbf{1}\big[\bar{\delta} \le \tau\big], \qquad \tau = 0.002. \quad (51)$$

The treatment removes both unsupported components, retains both supported ones, and completes withdrawal, recomputation, and rollback wherever it removes, reaching 1.000 on proposal, final account, useful removal, retention, and coherence against 0.500 for the append-only arm; mean private NMSE is 0.0052 against 0.0187, and the per-pair ratios show the point, since on the auxiliary task the arms differ by about one per cent in private-test MSE while disagreeing about whether the component belongs in the account at all.

Criteria for a successful removal. Across these records a removal counts as successful only when four recorded conditions hold jointly,

$$\mathrm{SuccRem}(c) = \mathbf{1}[D_c] \cdot \mathbf{1}[N_c] \cdot \mathbf{1}[R_c] \cdot \mathbf{1}[P_c] : \quad (52)$$

the challenged component is defeated by admissible counterevidence, every retained component remains necessary, dependent state is recomputed, and rollback provenance is complete. Predictive redundancy alone never satisfies the first factor. Under this count the component-revision treatment records six successful removals in six removal cells and zero unnecessary ones; the gated deletion arm records zero of either kind, consistent with its empty eligibility set; and the validation-only controller fails the second factor 34 times. The count is defined over pre-specified components and schedules, and certifies neither autonomous choice of what to challenge nor unrestricted falsification.

E.6 Per-task results and failure traces

This subsection describes what the frozen record holds at each level of granularity, records the three-run lineage of the two-anomalies suite with each run's main numbers, and narrates one evidence-to-decision trajectory from a single frozen confirmatory cell.

What the frozen record holds. Every experiment occupies its own release package with a frozen design, the settings, the modules, the runner and analyzer, deterministic tests, and a no-GPU preflight rehearsal; launches never overwrite each other. The unit a reader traces is the cell, keyed by task, seed, and condition, holding the complete recorded trajectory, its ordered decisions with raw generations and validation errors, its evidence rounds, its final state, and a completion marker carrying the digest of the record it certifies. The confirmatory two-anomalies run is split across two servers with disjoint task halves, merged only after both report complete with identical source digests; two post-completion fixes touched the analysis layer alone.

The two-anomalies run lineage. Three versions of the two-anomalies design were executed with frozen specifications. The first two are development lineage; the third is the final evaluation used in the main text. Each version fixed its specification before launch and recorded its outcome afterwards. Successive versions used fresh seeds because preceding trajectories had informed design changes. Fresh seeds prevent sample reuse but do not make the task distribution independent of development. Two gate definitions recur in Table 26. The comparative gate scores the false-retention delta over the six V-type cells of each arm,

$$\begin{aligned} \rho_V^a &= \tfrac{1}{6}\textstyle\sum_{c\in\mathcal{V}} \mathbf{1}\big[\text{arm } a \text{ retains the auxiliary account in cell } c\big], \\ \Delta_{\mathrm{FR}} &= \rho_V^{\mathrm{fit}} - \rho_V^{\mathrm{ctrl}} \;\geq\; 0.5, \end{aligned} \tag{53}$$

and the bounded-verification-cost subgate scores the six N-type pairs on actions to first commitment,

$$\frac{1}{6}\sum_{p\in\mathcal{P}_N} \mathbf{1}\big[a_p^{\mathrm{ctrl}} \leq a_p^{\mathrm{fit}} + 1\big] \;\geq\; \frac{5}{6}, \qquad a_p = B+1 = 4 \text{ for a never-committing trajectory,} \tag{54}$$

with shared budget $B = 3$ equal-cost evidence actions.

Run v1 failed the comparative gate at $\Delta_{\mathrm{FR}} = 0.333$. Reading all 30 trajectories located two defects: the 1024-token generation cap cut off long deliberations before a decision object appeared, with the resulting fallbacks concentrated in the policy arms, and the baseline paragraph failed its manipulation check, selecting the probe in 11 of 12 cells. Run v2 raised the cap to 4096 tokens and made the baseline's validation scope explicit; the construct then held and the comparative gate passed at 0.833, but the control paragraph named the separation precondition without a commit trigger, so two N cells overspent before committing, one V cell ended in a terminal no-commit, and the bounded-cost subgate fell to 4 of 6. Run v3 gave the control paragraph the same commit-trigger and remaining-budget clauses the fit paragraph carried, changing nothing else; both defects closed, every cell ended committed, the comparative gate passed at exactly the prespecified minimum of 0.500, and V dominance failed at 3 of 6. The wording line was declared closed at v3.

Table 26: The three frozen two-anomalies runs, each with 30 cells and 12 matched pairs plus a six-cell forced-probe component. Fallback cells count runs with at least one fallback decision; V dominance and N non-inferiority are pair counts on the mixed-regime private test. N retention uses non-fallback runs: the v3 value 5/5 excludes one N run with a fallback, while final outcomes in Table 27 cover all six N runs.

Run	Seeds, generation cap	Gate record	Fallback cells	Disposition
v1	4201, 4303, 4409; 1024 tokens	Validity PASS; control arm PASS (V rejection 6/6, N retention 6/6, bounded cost 5/6); comparative FAIL, $\Delta_{\mathrm{FR}} = 0.333$; private test FAIL (V dominance 2/6, N non-inferiority 6/6)	4 fit-guided, 3 forced-probe, 0 control	Development lineage; diagnosed as a 1024-token generation cap plus a failed baseline manipulation check
v2	5101, 5209, 5303; 4096 tokens	Validity PASS; comparative PASS, $\Delta_{\mathrm{FR}} = 0.833$; control arm FAIL on bounded cost alone (4/6) with V rejection 5/5 and N retention 5/5; private test FAIL (V dominance 4/6, N non-inferiority 6/6)	2 control, 0 fit-guided, 0 forced-probe	Development lineage; diagnosed as a missing commit trigger in the control paragraph, with one terminal no-commit
v3	6101, 6211, 6307; 4096 tokens	Validity PASS; control arm PASS (bounded cost 6/6, V rejection 6/6, N retention 5/5, probe before first commitment 12/12, re-diagnosis 6/6, zero terminal no-commits); comparative PASS, $\Delta_{\mathrm{FR}} = 0.500$; private test FAIL on V dominance 3/6 with N non-inferiority 6/6	1 control, 0 fit-guided, 0 forced-probe	Confirmatory record of Section 4.3

Table 27 gives the twelve matched pairs of the frozen v3 run. Across the arms, the control arm selects the out-of-regime probe in 12 of 12 cells and always before its first commitment, while the fit-guided arm selects it in 5 of 12; mean actions to first commitment are 1.583 for the control arm and 2.583 for the fit-guided arm; claim miscalibration is zero in both arms; and the forced-probe component, in which the probe outcome is injected before the first decision, records retention of the auxiliary account in 4 of its 6 cells, showing that access to the probe does not force selection of the model that transfers. The fit-guided instruction still prioritises in-scope validation, so this is not an isolated test of updating under a shared scientific objective.

Table 27: The twelve matched pairs of the frozen v3 two-anomalies run. False retention is scored on V-type cells only, where retaining the auxiliary account is the error; on the six N-type pairs both arms retain the genuine auxiliary readout. Here a is actions to first commitment and private-test NMSE is on the sealed mixed-regime panel; the three non-dominant V-type pairs are pairs in which the fit-guided arm also reached the correct structural account.

Task	Seed	Type	Fit false retention	a fit / control	Fit NMSE	Control NMSE
Alloy latent parent	6101	N	no	3 / 1	0.026143	0.026148
Alloy latent parent	6211	N	no	2 / 3	0.022818	0.023029
Alloy latent parent	6307	N	no	3 / 1	0.026251	0.025905
Bioprocess latent parent	6101	N	no	3 / 2	0.028832	0.028852
Bioprocess latent parent	6211	N	no	2 / 1	0.034623	0.034425
Bioprocess latent parent	6307	N	no	3 / 2	0.030376	0.030423
Thermal regime family	6101	V	yes	2 / 1	0.567989	0.015452
Thermal regime family	6211	V	no	3 / 2	0.020421	0.020572
Thermal regime family	6307	V	no	2 / 1	0.022126	0.022126
Transport relation family	6101	V	yes	3 / 1	0.388958	0.014374
Transport relation family	6211	V	no	2 / 2	0.015454	0.015454
Transport relation family	6307	V	yes	3 / 2	1.161603	0.013667

One evidence-to-decision trajectory. The strongest dominance pair, the transport task at seed 6307, is narrated here from its frozen control-arm cell. The design audit passes with an initial relative margin of 0.5701 and no leaked terms, and the offline block records candidate 2 as the auxiliary account and candidate 3 as correct. Table 28 gives the recorded sequence: the arm probes first, withholds commitment through two decisions, commits to the auxiliary account when two same-regime batches restore the in-regime ordering, and at the final decision, with the budget exhausted, reverses to the correct account at mechanism level, matching the ceiling the probe established.

Table 28: Recorded decision and reveal sequence of the frozen control-arm cell of the transport relation-family task at seed 6307. NMSE triples are per-candidate values from the facts-only appraisal, ordered candidate 1, candidate 2, candidate 3; the appraisal carries no recommendation.

Step	Recorded object	Content
Decision 1	Initial panel visible	Selects the one-shot out-of-regime probe; commitment withheld; no rejections; claim level predictive
Round 1	Probe outcome	Out-of-regime probe; coverage declared scope and a wider operating regime; NMSE 0.064381, 0.418749, 0.024307; claim ceiling raised to mechanism
Decision 2	Two evidence records visible	Selects a repeatable same-regime batch; commitment withheld; no rejections; claim level none
Round 2	Batch outcome	Same-regime batch, first occurrence; coverage declared scope only; NMSE 0.015070, 0.007808, 0.012536; claim ceiling predictive
Decision 3	Three evidence records visible	Selects the same-regime batch again; commits to candidate 2; rejects candidates 1 and 3; claim level predictive
Round 3	Batch outcome	Same-regime batch, second occurrence, taken with a standing commitment to candidate 2; NMSE 0.024047, 0.012058, 0.017438; claim ceiling predictive
Decision 4	Four evidence records visible; budget exhausted	No evidence action; commits to candidate 3; rejects candidate 2; claim level mechanism

Two recorded features matter. All four decisions parsed on the first generation with no fallback, and accept-auxiliary and reject-auxiliary both validate in either arm, so the reversal came from the policy under an unchanged validator. The endpoints keep the transient error, one evidence action taken while a false account stood, even though the cell ends final-correct; its sealed NMSE is 0.013667 against 1.161603 for its matched fit-guided partner, the largest of the three dominance factors.

Development runs and pooling. Runs v1 and v2 keep their own frozen packages, the superseded harness formulations stay as labelled lineage, the validation-only controller stays diagnostic, and only the final ordered-program iteration enters its pooled estimate; the inventory tables of Appendix E.1 carry the use label for every completed experiment.

F Prompts and reproducibility

This appendix reproduces what the model actually saw, including the two policy paragraphs that are the entire treatment of the two-anomalies comparison, and the digest machinery that pins what executed.

F.1 PROMPT AND SCHEMA INVENTORY

Every model call in the prototype loop and in the policy harnesses is issued from a frozen template with typed placeholders, so a run is determined by its configuration together with the templates recorded here. The production-loop templates are quoted from the versioned specification, the harness decision templates from the frozen source pinned by Eq. (55). Six templates cover the loop and the harnesses: the feature request and its repair, the semantic matcher, the hypothesis refinement, the optional value generator, and the harness decision object; each has one issuing stage and one failure route, and the complete inventory ships with the release.

Proposal contract. The LLMPROPOSE template instructs the model to act as an iterative scientific investigator. Its placeholders are filled, in fixed order, with the step index and step budget, the question q, the literature block for the current exposure of $\mathcal{K}_{\text{scope}}$, the ranked shortlist C_t from RANKCANDIDATES, the conditioning hypothesis, the measured set $\mathcal{O}$, the previously requested variables, the same-exposure exclusion set, and the forbidden list. The instruction block reads:

Proposal contract

```
You are an iterative scientific investigator.

You already have a current scientific hypothesis and a current set of measured variables.
Use the retrieved literature to decide the next currently unmeasured variable, or an allowed
bounded component, that would most improve, correct, or simplify the current hypothesis.

Prefer underlying variables over indirect bookkeeping fields.
You ARE allowed to request a new variable that is not currently measured.
The measured-variable list below is context, not a menu of allowed choices.
Do not request anything already measured.
Do not request the target variable itself.
Do not request anything in the forbidden list.
If the literature still points to hidden, latent, unresolved, proxy-only, or missing mechanism
terms, do NOT stop yet.  In that case, request a missing underlying variable or an allowed bounded
mechanism component.
Only stop when the current hypothesis is already sufficient and the literature no longer indicates
a missing mechanism term.

Return ONLY one of the following:
<feature>snake_case_variable_name</feature>
or, only if the current setting allows component commits:
<component>snake_case_variable_a, snake_case_variable_b</component>
or
<stop>I have enough information</stop>
```

The shortlist block is introduced by the line `Literature-supported candidate shortlist (ranked best-first)` and followed by the instruction to prefer one of its entries while it remains consistent with the current hypothesis gap; the template closes by requiring a single XML block with no extra explanation. Under the explicit component-commit condition the `<component>` alternative is enabled, so one call may request a small bounded mechanism component, for example two variables that the same exposure presents as inseparable.

Proposal repair and stop repair. Raw output is parsed through a fixed ladder from the stop tag down to a short identifier-like phrase; when nothing survives, the repair template re-prompts with the reason and the original context, and the same template repairs premature stops when the evidence-gap guard fires.

Semantic matching. The optional matcher template restricts matches to close rephrasings and directs a NONE for neighbouring concepts, correlates, and proxies; its inventory excludes every role label, with the resolution order and confidence threshold in Appendix D.4.

DSL generation and repair. The refinement template instructs the model to refine the current DSL hypothesis from the literature and the measured variables alone, to include a newly measured variable when it is causally useful, and to explain any omission; it returns one `<dsl>` block assigning the target plus an `<answer>` block, and failing candidates are repaired against the validator errors. Under the multi-prior setting, the call is repeated and the best validation refit is kept.

Decision template of the policy harnesses. The harness issues one merged decision template per decision point. In the two-anomalies cells there is one decision before each of the three evidence actions and one final decision after the last reveal; in the forced-probe cells the injected probe consumes

one budget unit, leaving three decision points. The template is assembled from blocks in fixed order: a shared preamble; SCIENTIFIC QUESTION AND SCOPE, the public task rendered as indented JSON with sorted keys; LIVE EXECUTABLE CANDIDATES, the frozen three-candidate bundle; STANDING COMMITMENT; VISIBLE TYPED EVIDENCE, everything revealed so far; REMAINING EVIDENCE-ACTION MENU with the remaining budget out of the shared three-action total; a single-paragraph INQUIRY POLICY; and the JSON schema. The policy paragraph is the only text that differs between arms. The shared preamble reads:

Shared decision preamble

```
You are one step in an iterative scientific inquiry with a fixed evidence budget. Use only the
material printed below. Ground truth, future evidence, and the private-test outcome are
unavailable. Repeatable actions may be selected again; each selection returns a fresh result and
costs one budget unit. Your commitment can change at any decision point; the final decision point
fixes it.
```

The fit-guided policy paragraph, verbatim:

Fit-guided policy paragraph

```
INQUIRY POLICY
Validation for this programme is measured within the declared operating scope. Prefer evidence
actions expected to improve the in-scope validation of the account the current evidence favours
most, and commit when its in-scope validation improves. Continue confirmatory observation of the
committed account with the remaining budget.
```

The epistemic-control policy paragraph, verbatim:

Epistemic-control policy paragraph

```
INQUIRY POLICY
Before committing, choose the evidence action expected to distinguish whether the auxiliary
readout is a genuine input to the mechanism or merely tracks the target within the current
operating scope. Once that separation has been observed, commit to the account that best survives
it, and continue confirmatory observation of the committed account with the remaining budget.
```

Once the budget is exhausted the menu block is replaced by the heading REMAINING EVIDENCE-ACTION MENU followed by the instruction that the budget is exhausted, that `evidence_action_id` must be set to `"none"`, and that the final decision must be made.

Merged action object. Each decision returns one JSON object covering evidence choice and epistemic state together, so an action and a commitment cannot be recorded under different beliefs. The schema block printed to the model reads, with the action rule and the example candidate identifier interpolated at runtime:

Decision object schema

```
Return exactly one JSON object and no prose:
{
  "evidence_action_id": exactly one action_id from the remaining menu,
  "commitment": "no_commit" or "commit_<candidate_id>",
  "rejected_candidates": ["candidate ids you now rule out; may be empty"],
  "claim_level": "one of: none, predictive, variable, mechanism, structure",
  "claim": "one sentence stating your current scientific claim"
}
```

The validators accept only menu actions and live candidate identifiers, reject a causal claim level as outside the declared scope, and never compare the object against the appraisal, so accepting and rejecting the auxiliary account both validate in either arm. Parsing allows two attempts; after the final failure the harness records a type-neutral fallback object, counted against the per-arm cap frozen in the design. (Appendix F.2).

F.2 CONFIGURATIONS AND EXPERIMENT MANIFESTS

Every confirmatory experiment is fixed before launch by a frozen design that declares the task list, the seeds, the model profile, the sample sizes, the inquiry budget, the pairing contract, the endpoint list, and the complete gate thresholds. Table 29 enumerates the registers of the six frozen confirmatory

designs behind Sections 4.3 and 4.2. A collision audit over every frozen design confirms that no seed triple recurs (Appendix D.3).

Table 29: Frozen design registers of the six confirmatory experiments. Cell counts come from the frozen evidence index of the matched-trajectory suite and, for the two-anomalies row, from its prespecified gate block; arms are the paired conditions, with the two-anomalies row adding a single-arm forced-probe component on its two V-type tasks. The cap is the maximum number of new tokens a decision call may emit, and panel sizes are training / validation / private-test rows per cell.

Experiment	Cap	Seeds	Cells	Matrix and panels
The matched-evidence comparison	1024	1301, 1409, 1511	24	4 tasks × 3 seeds × 2 arms; 192 / 192 / 512
The matched-menu comparison	1024	401, 503, 607	24	4 tasks × 3 seeds × 2 arms; 192 / 192 / 512; action budget 2
The reopening comparison	1024	1601, 1709, 1811	12	2 tasks × 3 seeds × 2 arms; 192 / 192 / 512
The diagnosis-stress comparison	2048	3301, 3407, 3511	24	4 tasks × 3 seeds × 2 arms; no numeric panels; 3 parse attempts
The component-revision comparison	1024	2503, 2609, 2711	24	4 tasks × 3 seeds × 2 arms; 96 / 96 / 192
The two-anomalies comparison	4096	6101, 6211, 6307	30	4 tasks × 3 seeds × 2 arms, plus 6 forced-probe cells; 192 / 192 / 512; action budget 3

The two-anomalies design fixes two N-type tasks, the alloy and bioprocess latent-parent tasks, and two V-type tasks, the piecewise thermal-response task and the transport relation-family task; an inquiry block with temperature 0.0, at most 2 parse attempts, a budget of 3 equal-cost evidence actions, and an adequacy bound $\mathrm{NMSE} \leq 0.15$ on the in-regime panel; and a confirmatory gate block holding 30 expected cells split into 24 main-matrix and 6 forced-probe cells, a matched-input fraction of 1.0, a per-arm fallback cap of 0.25, the control-arm subgates, the comparative false-retention delta minimum of 0.5, and the private-test margins of 0.05 for V-dominance on at least 5 of 6 pairs and for N-noninferiority on 6 of 6 pairs.

Launch record and resume checks. At launch the runner writes a manifest holding the experiment identity, the model profile, the frozen cell list, and a digest map over the eleven behavioural source modules,

$$H_{\mathrm{src}} = \left\{ \big(p,\ \mathrm{sha256}(\mathrm{bytes}(p))\big) : p \in \mathcal{P} \right\}, \qquad |\mathcal{P}| = 11, \tag{55}$$

covering the design, configuration, profile, benchmark, policy, calibration artifact, runner, analyzer, and shared runtime. Every resume re-verifies the cell matrix, the profile, and the digest map against the manifest and aborts on any mismatch, so a run cannot continue under edited code or an edited design; the matched-trajectory runners enforce the same three checks.

Calibration freeze. The generator-level calibration of the two-anomalies suite is pinned twice, by its file digest inside Eq. (55) and by a content hash over its canonical serialisation,

$$h_{\mathrm{cal}} = \mathrm{sha256}\big(\mathrm{json}_{\mathrm{canon}}(F \setminus \{f_{\mathrm{cal}}\})\big), \tag{56}$$

recomputed at every launch and re-checked when the two server halves merge. The registered quantities are computed on 100,000-row populations and never on run panels, so no sample-level gate can select seeds; the frozen bands, margins, and the $K = 200$ indistinguishability audit at accuracy 0.601 inside the chance band $[0.35, 0.65]$ are stated in Appendix C.4.

Corpus versions. Literature exposure runs over frozen document pools, so the corpus a run consumes is fixed by a dated benchmark version whose specification, quality report, task specifications, and every document are pinned by digest (Appendix C.3).

The accepted record. A final evidence index names the five accepted matched-trajectory blocks and fixes fifteen digests over their records; the two-anomalies record is the two per-server runs and their merge, whose validity gate reports 30 of 30 cells, matched inputs at 1.0, consistent digests across hosts, and per-arm fallback rates inside the 0.25 cap.

F.3 Models, compute, and execution cost

Model and decoding. Main comparisons run DeepSeek-R1-Distill-Qwen-32B from a local checkpoint at temperature 0; the three model profiles differ only in the generation cap, 1024, 2048, or 4096 new tokens, and the cap is part of the measured design (Appendix D.3).

Execution harness. Each host runs one worker per GPU, pinned to its device, with both arms of a pair executed inside the same worker and one loaded model instance.

Per-cell cost. The 30 confirmatory two-anomalies cells issued 117 model calls, three of them repairs, and summed to 1.49 hours of single-device time (Table 30).

Table 30: Execution cost of the frozen confirmatory two-anomalies run, computed from the merged per-cell metrics and from the resource record of each cell, which holds its call count, elapsed time, and prompt and output volume. Elapsed time is wall clock per cell on one pinned device, and character counts are summed over all calls of a cell.

Condition	Cells	Mean model calls	Mean elapsed (s)	Mean prompt characters	Mean output characters
fit-guided policy	12	4.08	189.6	31,057.7	11,788.2
epistemic control	12	4.17	190.1	34,658.5	14,985.2
fit-guided, forced probe	6	3.00	137.0	23,355.0	10,001.3
all cells	30	3.90	179.3	30,957.5	12,709.6

Software environment. The frozen pipeline records its interpreter and package state alongside the source snapshot: CPython 3.13.13 on Linux, with NumPy 2.2.6, SciPy 1.14.1, and Matplotlib 3.10.9 among the pinned packages.

F.4 Artifact lineage and regeneration

Source pinning. A revision identifier alone underdetermines the executed bytes, so the per-file digest map of Eq. (55) is the operative lock; the merged two-anomalies report repeats the digest map of both servers, and its cross-record consistency check passed, establishing that the two halves executed byte-identical sources. The same rule pins the shared client and orchestration modules and every benchmark document, so an edit anywhere in the behavioural path invalidates a resume instead of silently changing a run.

Figure regeneration. Every result figure is produced by one idempotent script that reads only the frozen records of its evidence block and recomputes every plotted value; the matched-trajectory summary script additionally asserts its recomputed endpoints against the frozen evidence index before drawing.

Exclusion of archived artifacts. Three mechanisms enforce the no-pooling rule in practice: a manifest that refuses to resume across a source or design change, an analyzer that refuses to report until every expected cell carries a digest-checked completion marker, and figure scripts that read the accepted records by explicit reference.

Superseded formulations and their diagnoses. Each superseded formulation carries its own frozen design, a written audit of the failure, and the analyzer's gate report, and each holds a seed triple no accepted block reuses. Table 31 collects the four defects and remedies; the full audits, including replay evidence that the failed contracts and never the scientific decisions caused the fallbacks, ship with the release.

G Open problems

The controlled studies leave the central problem of autonomous representation revision unresolved. Following the concrete-problem approach of McAllister et al. (2017), we describe what further evidence would distinguish progress from successful execution of a supplied policy.

Choosing the level of revision. An agent must distinguish an inadequate candidate, insufficient search, missing measurements, model-class limitations, and uncertainty that warrants replication or abstention. The current experiments supply important parts of this structure through catalogues, candidates, and evidence schedules. A stronger test would withhold the failure label, give all methods the same scientific objective and resource limits, and evaluate whether the agent chooses an appropriate response, including leaving the representation unchanged. Finite observations do not generally prove inexpressibility; the target is justified, revisable diagnosis under explicit assumptions.

Table 31: The four superseded formulations excluded from the confirmatory estimates of the matched-trajectory suite, with the completed matrix of each, the gate or construct that rejected it, the audited cause, and the change its successor made. Status gives passed gates over prespecified gates together with the completed matrix and its seed triple; rates are read from each run's frozen acceptance record. Every successor draws a fresh seed triple, and none of these runs is pooled with the accepted blocks of Table 12.

Formulation	Status and matrix	Defect	Remedy	Superseded by
Matched evidence, first	FAIL, 6/8 gates; 24 cells, 12 pairs; seeds 101, 211, 307	`diagnosed_gap` conflated the unresolved gap with the already repaired deficiency, so family response scored 0.000 in both families against a 0.75 minimum; the trajectory-improvement disjunction tied on all three of its terms; one infinite private-test ratio was dropped as non-finite instead of counted as a failure	Split the field into `current_gap`, `identified_gap`, and `resolution_status`; count infinite ratios as failures; supply both arms one deterministic level-aware appraisal; prespecify retirement, typed repair, and bounded claim release	the second formulation
Matched evidence, second	FAIL, 12/13 gates; 24 cells, 12 pairs; seeds 701, 809, 907	Treatment fallback rate 1.000 against a 0.25 cap: the schema encoded one action redundantly across four fields; 11 of 14 fallback rounds already carried the correct core action, and all 14 replay valid under the successor contract	Narrow the controller contract to a five-field scientific action and move state bookkeeping to a deterministic reducer, with no gate, evidence ceiling, or isolation requirement weakened	the accepted comparison
Reopening, first	FAIL, 10/11 gates; 12 cells, 6 pairs; seeds 1009, 1103, 1201	Fallback rate 1.000 in both arms: the prompt exposed a ledger naming panel alpha while the validator admitted only panel beta onward; 23 of 24 validation errors were invalid evidence identifiers	Give cumulative provenance and the licensed transition to a shared deterministic reducer, and reduce the controller to event response, selected candidate, claim level, and claim text, with the benchmark and all eleven gates preserved	the accepted comparison
Diagnosis stress, first	FAIL, 10/14 checks; 24 cells, 12 pairs; seeds 1901, 2003, 2111	`diagnosed_gap` mixed the initial limitation with the residual state left after the response, so correct repairs were rejected and treatment diagnosis and response accuracy both recorded 0.583 (Appendix E.2)	Request `initial_limitation` in four generic classes, drop the diagnosis-improvement gate for a declared scientific reason, and tighten the surviving thresholds	the accepted comparison

Acquiring concepts and measurements beyond a catalogue. The prototype selects measurements from quantities defined in advance. Broader discovery requires proposing new scientific distinctions and connecting them to a feasible measurement or test. An inadequate measurement interface can itself need revision; adding an instrument or intervention is distinct from adding another hypothesis over existing data.

Mechanisms, interventions, and identification. Ordered-program recovery under measured intermediates does not establish identification of arbitrary causal structure. Observational identification may be possible under additional assumptions; observationally equivalent candidates require further assumptions or evidence that separates them (Pearl, 2009; Choo et al., 2023). Intervention interfaces should specify what is manipulated, how outcomes are generated, and the assumptions supporting the estimand. Simulation can reveal a candidate's consequences, but agreement with a simulator is not independent confirmation of its assumptions about the world. Claim-specific support must track the candidate, the regime, how the evidence was obtained, and possible defeating evidence; the current claim-ceiling check enforces only a narrower benchmark rule.

Exploration and generalisation. Discriminating current hypotheses does not exhaust scientific inquiry. Replication, measurement checks, and exploratory observations can expose limitations shared by every live candidate (Stanford, 2006). Future evaluation should include task families not used in policy development and stronger alternative design and revision methods. Prospective scientific use would additionally require independently checked predictions and evidence of real-world mechanism discovery. These are open directions; the current studies do not establish them as outcomes.

H STUDIES OUTSIDE THE MAIN CLAIMS

Two supplementary tracks use designs that differ from the matched comparisons in Section 4. We report them separately; the main text cites one descriptive result from the proxy study, without pooling its cells with the matched experiments. Each states what it asks, what it is entitled to conclude, and why it is scored separately, so that a reader can tell which conclusions of the paper do not depend on it.

H.1 PROXY VARIABLES AND MEASUREMENT SUBSTITUTES

A proxy is an informative measurement of another quantity, so it can improve prediction while remaining the wrong scientific input to the target mechanism. The proxy-contamination track tests what the proxy-free panels cannot expose: whether the discovery loop keeps a measurement substitute in its final explanation after the source variable has been recovered, and whether a declared typed replacement policy removes that substitute safely. Three declared reasons keep it outside: Stage A alters the initial measured set, so its cells are unmatched against the proxy-free arms; Stage B consumes a declared measurement relation the loop does not infer, so its result is policy feasibility under supplied metadata; and the predictive endpoint pools targets whose scales differ by orders of magnitude, so the pooled mean supports no prediction claim either way. Section 4.5 draws one number from this track, the spurious-structure inclusion rate (structural contamination in the record) 0.031 against 0.184.

A declared relation becomes actionable only after the source is measured. A proxy P of a source Z is generated as a noisy affine measurement,

$$P = \alpha Z + \eta, \qquad \eta \sim \mathcal{N}\big(0, (\kappa \cdot \mathrm{std}(Z))^2\big), \tag{57}$$

so it carries real information and can enter a validation-selected basis on predictive grounds alone. The optional replacement policy operates on that declared relation; a replacement must name its substitute, pass resampled non-inferiority, retain the source, remove at most one candidate per recomputation, and write a full ledger entry with rollback state.

Stage A: initial proxy exposure contaminates the final structure. Stage A runs the discovery loop over all ten reference scenarios and three seeds under two conditions, no proxy visible at the start against both proxy columns visible in $\mathcal{O}_0$, with everything else held fixed within a pair; the endpoint is the share of proxy columns among the final account's non-target variables.

Initial visibility raises the runs that retain a proxy from 4 of 30 to 20 of 30 and the retained proxy terms from 5 to 25, so the mean proxy rate grows almost sixfold, from 0.031 to 0.184, and nine of the ten scenarios contribute at least one contaminated run. Hidden recall falls from 1.000 to 0.867 and final variable precision from 0.969 to 0.802, while hidden precision moves the other way, from 0.950 to 0.967. Strict proxy-free recovery falls from 26 to 10 of 30 cells, so discovering the source does not by itself remove the substitute. The predictive endpoints disagree in direction, the mean private-test error rising from 35.253 to 38.144 while the median falls from 1.762 to 1.671, so this track supports no prediction claim either way. Stage A records no typed proxy-to-source-to-removal event, the gap the isolated Stage-B track addresses.

Stage B: the typed policy removes the substitute in every cell. Stage B isolates the policy from the discovery loop on two controlled measurement-replacement scenarios and three seeds, giving six cells per method, one replacing a texture index by grain size and one an imaging score by microcrack density. The append-only control performs no replacement; the typed sequential policy sees the declared relation, records the lifecycle events, and freezes its decision before private-test metrics are released under a one-shot lock, so a released value cannot re-enter its own decision. The typed method removes the mapped proxy in 6 of 6 cells and the append-only control in none, both retain the named source and complete the lifecycle in every cell, and the append-only column set keeps the proxy and an unsupported noise term while the typed set keeps context and source.

Two properties bound the claim. Every validation delta is positive while every resampled stability interval has a negative lower bound, so acceptance rests on the declared relation together with resampled non-inferiority instead of a demonstrated validation gain. The typed method gives the lower private-test MSE in five of the six cells, the exception being the grain scenario at seed 1, where it reads 0.21423 against 0.21395. Stage B therefore establishes policy feasibility under declared measurement metadata, and it establishes neither autonomous proxy recognition nor end-to-end replacement inside Algorithm 1: the typed relation is supplied by the task, and each scenario is constructed so that exactly one mapping is live.

H.2 The real-data feature-selection analogue

Symbolic refit searches a sparse additive class of univariate basis terms over the measured variables (Section 3.1),

$$\widehat{Y} = \beta_0 + \sum_j \sum_m \beta_{jm} \phi_m(X_j). \tag{58}$$

This track asks whether Eq. (58) has a real-data analogue outside the synthetic SCM panels: a workflow that starts from a small measured marker set, considers further candidate markers, and estimates their additive contributions to a held-out endpoint. The study is a controlled analogue of the discovery loop, in which a hand-specified exposure schedule and a ridge selector stand in for RETRIEVE, RANKCANDIDATES, LLMPROPOSE, GENERATEANDVALIDATEDSL, and SYMBOLICREFIT, with $\phi_m(X_j) = X_j$ throughout, so every basis term is one standardised marker.

Dataset, endpoint, and repeated splits. The data are the UCI Wisconsin Diagnostic Breast Cancer table: 569 samples, 212 malignant and 357 benign, with 30 cell-nucleus morphology descriptors treated as predictive markers with no assumed causal role, and $Y = 1$ for malignant. Three seed-derived permutations of the table are each split 341/114/114 into training, validation, and sealed test, with columns standardised from training statistics alone; these are repeated splits of one cohort and no independent validation.

Exposure schedule and commit rule. Four fixed markers start every run, and the remaining 26 columns are exposed in five declared families; at each step every candidate of the current family is added separately, refit by ridge under one fixed penalty, and the best is committed when it lowers validation MSE by more than 10^{-6}, at most one per family. Every family clears that threshold in every seed, so all progressive models hold nine features, a size fixed by the schedule.

Conditions and controlled comparisons. Four conditions share the split, standardisation, and estimator within a seed (Table 32): the initial four markers, progressive discovery, a size-matched random subset that separates the selection rule from the feature count, and an all-feature reference that carries no upper-bound guarantee under a fixed penalty.

Table 32: Endpoints of the real-data feature-selection analogue over three repeated splits of one cohort. Per-split values are sealed private-test MSE, the mean columns carry the sample standard deviation of the three splits, and AUC is the mean sealed private-test area under the ROC curve.

Condition	Feat.	Split 0	Split 1	Split 2	Mean MSE	Mean AUC
Initial only	4	0.0830	0.0886	0.0635	0.0784 ± 0.0132	0.9883 ± 0.0065
Progressive discovery	9	0.0650	0.0622	0.0629	0.0634 ± 0.0015	0.9930 ± 0.0052
Random matched	9	0.0744	0.0715	0.0762	0.0740 ± 0.0024	0.9896 ± 0.0027
All-feature reference	30	0.0604	0.0681	0.0566	0.0617 ± 0.0059	0.9849 ± 0.0060

Progressive selection lowers mean private-test MSE from 0.0784 to 0.0634, about 19 percent, and it wins on all three splits against the initial set and the size-matched random subset at 0.0740 (Table 32). That count-matched control improves the baseline by only 0.0044 on average and loses to it on split 2, so five arbitrary columns do not account for the gain. The margin over the initial set is uneven, 0.0264 on split 1 and 0.0006 on split 2. The progressive model does not beat the all-feature reference at 0.0617, trailing it by 0.0017 in mean private-test MSE and winning on one of the three splits, so the schedule recovers most of the available signal with under a third of the descriptors. Ranking is close to saturated in every condition, so the initial markers already rank well.

Selected-set stability. Pairwise Jaccard similarity of the five discovered markers is 0.111, 0.667, and 0.250 across the three splits, mean 0.343, so the predictive gain is more stable than the identity of the markers producing it; the one marker selected everywhere is partly a schedule artifact, and the rest tracks the table's correlation structure, where mean radius and mean perimeter correlate at 0.998.

Scope and protocol limits. The table carries no causal graph and no interventional action, so every claim sits at the observational-association rung, and the hand-authored schedule and selector exercise none of the retrieval, proposal, or DSL machinery. The run also serialises sealed metrics for every candidate trial, which falls short of the release-once rule even though no automated decision consumed them, and three seeds support no publication-grade uncertainty statement.