BEYOND THE BEST AVAILABLE CAUSE: CALIBRATED VERIFICATION AND TRIAGE UNDER HYPOTHESIS-SPACE INADEQUACY

Anonymous authors
Paper under double-blind review

ABSTRACT

From clinical practice to LLM reasoning, causal diagnosis commits to the top-ranked candidate from a set of explanations. When the true cause is absent from that set, even the best candidate can be wrong. We call the absence of any task-compatible candidate hypothesis-space inadequacy, and ranking alone cannot detect it. Diagnosing it requires separating it from *selection failure*, where an adequate space still yields a wrong pick. We propose **C**ausal **A**dequacy **V**erification with **E**vidence-guided **A**ction and **T**riage (CAVEAT) with a dual-null test design, one test and one repair for each failure type. The core idea is to make inadequacy testable without knowing the true cause, by defining compatibility through the consequences the task depends on and comparing every candidate's predictions with the observed evidence. The first test assesses support for the space and requests revision when support is insufficient; the second checks the selected candidate and prompts distinguishing evidence when the choice remains unresolved. Among task-incompatible selections, calibration limits how often commitment occurs, assuming past and new failures are exchangeable under the same workflow. On a clinical benchmark, retrieved candidate lists omit the recorded diagnosis in 78% of cases. Across physical, biological, and clinical tasks, CAVEAT reduces wrong commitments, with retained usefulness depending on the evidence. In controlled LLM reasoning, test-guided repair yields more correct commitments than either fixed repair, even with imperfect candidate supply.

1 INTRODUCTION

Causal diagnosis, root cause analysis, and LLM reasoning share a common pattern. They generate an active hypothesis space and select the candidate that ranks highest. When none of the candidates explains the incident sufficiently well for the downstream task, a closed-set system can still commit to its selection without detecting the problem or searching for a possibly better candidate outside the active hypothesis space. We call this problem *hypothesis-space inadequacy*. A similar problem appears in medical premature closure, in which a diagnostic conclusion is made without due consideration of reasonable alternatives. A retrospective study of 100 diagnostic error cases found that 90 led to patient harm, including 33 fatalities, and premature closure was the most common cognitive contributor (Graber et al., 2005). Evaluation can also hide limits of the available evidence. OpenRCA excludes records without clear anomalies at the root-cause component, leaving performance on those cases unmeasured (Xu et al., 2025). OpenRCA 2.0 shows that a system can identify the correct cause without verifying how it produced the observed failure (Fang et al., 2026). The same issue appears with LLM proposers. On our controlled causal task, self-reflection does not resolve the problem, and raw errors remain frequent across 3B–32B models (Section 5), consistent with reported limits of self-correction without external feedback (Huang et al., 2024). Hence, improving the ranking model alone is not enough. We first need to ask whether the hypothesis space is inadequate, because the answer determines whether to revise the candidates or seek distinguishing evidence before committing. We call this process *adequacy-aware causal diagnosis* and show that the question can be tested without knowing the true cause.

Existing methods solve parts of the problem. The first part is constructing the space. Causal discovery learns structures within a specified model class (Pearl, 2009; Spirtes et al., 2000), while

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

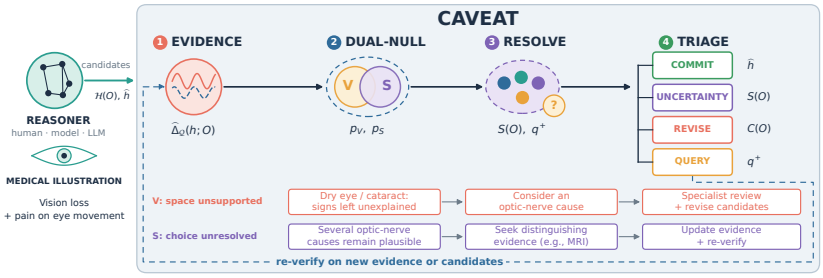

Figure 1: **CAVEAT: verify the space, then resolve the choice.** Two calibrated gates test hypothesis-space inadequacy (V) and selection failure (S) to guide repair. Bottom: a hypothetical visual-loss case (MedlinePlus Medical Encyclopedia, 2026).

latent-variable and LLM-based methods propose new variables and graphs (Dong et al., 2024; Li et al., 2024; Liu et al., 2024; Abdulaal et al., 2024). Constructing these explanations does not by itself establish their suitability for the downstream task. A second part is refusing or qualifying a prediction. Open-set recognition, selective prediction, and conformal methods provide rejection rules or uncertainty sets (Chow, 1970; Bendale & Boult, 2016; Hendrycks & Gimpel, 2017; Geifman & El-Yaniv, 2019; Vovk et al., 2005; Angelopoulos & Bates, 2023), and recent variants account for unseen labels (Xie et al., 2025), the removal of unsupported claims (Mohri & Hashimoto, 2024), and budgeted evidence acquisition (Xu et al., 2026). Closest to our setting, Li et al. (2026) study coverage limits when sampled answer sets contain no semantically acceptable response. These methods address missing answers and alternatives to commitment, whereas we distinguish a missing task-compatible explanation from a wrong selection within an adequate space. A third part is checking a given structure. Causal falsification tests structures or effects against observations (Faller et al., 2024; Su & Henckel, 2022; Orchard et al., 2026), sensitivity analysis examines vulnerability to unmeasured confounding (Cinelli & Hazlett, 2020), and posterior predictive checks assess model fit (Gelman et al., 1996). Rejecting one candidate, however, does not establish inadequacy of the whole space. The last part is deciding under uncertainty. Learning-to-defer methods hand difficult cases to an expert (Mozannar & Sontag, 2020), conformal risk control bounds expected losses (Angelopoulos et al., 2024), and active intervention design distinguishes graphs within an equivalence class (Hauser & Bühlmann, 2014). Our focus is the combination: testing adequacy through task-relevant consequences, separating it from selection failure, and using the two results to guide repair.

The new object of verification is therefore the adequacy of a hypothesis space for a task. A missing compatible explanation requires revising the space, and an unresolved choice calls for checking the selection or seeking distinguishing evidence. We present CAVEAT as a verification and triage layer that connects these two checks to commitment and repair. CAVEAT uses two gates for hypothesis-space inadequacy and selection failure, based on discrepancies between observations and candidate predictions on task-relevant consequences. The space gate tests inadequacy using the best candidate's evidence discrepancy; the selection gate then evaluates the proposer's choice. Both gates use conformal calibration on past incidents of the corresponding failure type, exchangeable with new incidents under the same workflow. We use standard conformal calibration to bound how often an incompatible selection receives commitment. We also characterize how candidate selection and evidence limit correct commitment, and derive a sufficient condition under which test-guided repair yields more correct commitments than always revising or always querying. Neither gate requires access to the proposer's internals or knowledge of the new incident's true cause. The gates also guide the next step: COMMITMENT, space REVISION, an UNCERTAINTY SET, or a QUERY for distinguishing evidence.

Contributions. (i) We formalize task-relative hypothesis-space inadequacy, make it testable through predicted and observed consequences, and distinguish it from selection failure. (ii) We introduce CAVEAT, whose separately calibrated gates control commitment conditional on task-incompatible selection and guide repair. We derive a sufficient condition for routing to outperform both fixed repairs. (iii) In primary real-data comparisons, forced choice is wrong on 49–94% of units while verification reduces wrong commitment. In controlled LLM reasoning, routing recovers more correct answers than either fixed repair, with gains persisting under imperfect candidate supply but varying across reasoners.

2 PROBLEM SETUP: HYPOTHESIS-SPACE INADEQUACY

Causal diagnosis supplies the explanation that looks better than the other candidates, while we ask whether any candidate could serve the real-world task. We first define the factors connecting explanation to downstream decision. For example, we call **incident** the case to be diagnosed, such as a patient presentation or a system failure; We then denote observed evidence O and an unknown **operative mechanism** $h^\star$, the true mechanism that generates the observation.

Causal diagnosis works as an upstream proposer G that constructs the **active hypothesis space** $\mathcal{H}(O) = \{h_1, \ldots, h_M\}$ based on observation, and selects $\widehat{h} \in \mathcal{H}(O)$, so $G(O) = (\mathcal{H}(O), \widehat{h})$. We do not assume that either $h^\star$ or a task-compatible alternative must be represented.

The **task**, choosing a treatment for example, is the decision to be made. It determines the **causal queries** $\mathcal{Q} = \{q_1, \ldots, q_K\}$, which are the observations that decision relies on, such as the pain on eye movement, the appearance of the optic disc, and the MRI finding in the visual-loss case of Figure 1. For each query q_k, a candidate h supplies a **predicted consequence** distribution $P_{h,q_k}(\cdot \mid O)$ directly or through a model, which would be compared against the **actual consequence** distribution $P^\star_{q_k}(\cdot \mid O)$ from the operative mechanism. We define this comparison as **discrepancy** d_k, measuring their gap: $r_k(h; O) = d_k(P_{h,q_k}(\cdot \mid O), P^\star_{q_k}(\cdot \mid O))$, normalized across queries.

Definition 1 (Task compatibility and hypothesis-space adequacy)**.** *With a tolerance $\epsilon \geq 0$ fixed, a candidate is **task-compatible** if its **worst-query discrepancy** is within tolerance; the space is **adequate** if it contains at least one such candidate:*

$$\begin{aligned} &\Delta_{\mathcal{Q}}(h; O) := \max_{k=1,\ldots,K} r_k(h; O), \\ &\mathcal{A}_\epsilon(O, \mathcal{Q}) := \{h \in \mathcal{H}(O) : \Delta_{\mathcal{Q}}(h; O) \leq \epsilon\}, \qquad \textit{adequate} \iff \mathcal{A}_\epsilon(O, \mathcal{Q}) \neq \varnothing. \end{aligned} \tag{1}$$

The tolerance ϵ, determined by domain requirements or independent development data, is the maximum threshold compared with a candidate's worst-query discrepancy.

Two failures, different responses. Two null hypotheses, indexed by V and S, identify the failures: 1) hypothesis-space inadequacy, $H_0^V : \mathcal{A}_\epsilon(O, \mathcal{Q}) = \varnothing$; 2) selection failure, $H_0^S : \mathcal{A}_\epsilon(O, \mathcal{Q}) \neq \varnothing,\ \widehat{h} \notin \mathcal{A}_\epsilon(O, \mathcal{Q})$. Section 3 calibrates a gate for each and requests a repair to each: revision for the first, further evidence for the second.

3 CAVEAT: EVIDENCE, VERIFICATION, AND TRIAGE

In an archived case from our controlled CLadder extension (Section 5; Figure 5), a language model selects one of four supplied graphs, none of which is task-compatible. CAVEAT scores each graph against the probed effects (Stage I). Its space gate rejects the space (Stage II), and it then requests a prespecified revision to expand the space to eight candidates together with a follow-up probe contradicting the original prediction (Stage III). Afterwards, the model's new selection passes reverification and is committed (Stage IV). Figure 1 shows all stages.

3.1 STAGE I: SCORE AGREEMENT WITH INCIDENT EVIDENCE

Imagine an early COVID-19 case assessed against a list of familiar illnesses: the best match could still miss the cause. As in the CLadder incident, ranking cannot reveal that all candidates fit poorly. We start by measuring agreement with evidence for the space and the selected candidate.

From evidence to discrepancy. Candidate h predicts the consequence distribution $P_{h,q_k}(\cdot \mid O)$ for query q_k. The corresponding distribution under the operative mechanism, $P^\star_{q_k}(\cdot \mid O)$, is unknown, so we estimate it from available evidence as $\widehat{P}^\star_{q_k}(\cdot \mid O)$. Repeated intervention measurements can supply this estimate directly. Replacing the unknown distribution in Section 2 gives $\widehat{\Delta}_{\mathcal{Q}}(h; O) = \max_k d_k(P_{h,q_k}(\cdot \mid O), \widehat{P}^\star_{q_k}(\cdot \mid O))$. Smaller discrepancies mean closer agreement (consistency conditions: Appendix B.2).

Two scores from the same discrepancies. The space needs one score and the selection another: A_V is the discrepancy of the best-fitting candidate, so a large A_V means nothing in the space fits,

and A_S compares the proposer's choice $\widehat{h}$ with its best alternative:

$$A_V(O) := \min_{h \in \mathcal{H}(O)} \widehat{\Delta}_{\mathcal{Q}}(h; O), \tag{2}$$

$$A_S(O) := \widehat{\Delta}_{\mathcal{Q}}(\widehat{h}; O) - \min_{h \in \mathcal{H}(O) \setminus \{\widehat{h}\}} \widehat{\Delta}_{\mathcal{Q}}(h; O). \tag{3}$$

For $A_S < 0, = 0, > 0$, the selection fits better, ties, or fits worse than its best alternative. A negative margin alone cannot prove a candidate is compatible, since a large A_V could still make it a poor fit.

3.2 Stage II: dual-null calibration

A patient with COVID can look like a patient with flu, so matching flu alone does not establish the diagnosis. Limited or noisy evidence can give an incompatible candidate a favorable score. Calibration controls the risk of accepting such failures.

Calibration rule. Let $\{O_i^V\}_{i=1}^{n_V}$ and $\{O_\ell^S\}_{\ell=1}^{n_S}$ be test-disjoint calibration incidents labeled as hypothesis-space inadequacy and selection failure. The lower-tail ranks count how often each failure type looked at least as supportive as the new incident:

$$p_j(O) := \frac{1 + \sum_{i=1}^{n_j} \mathbf{1}\{A_j(O_i^j) \le A_j(O)\}}{n_j + 1}, \qquad j \in \{V, S\}. \tag{4}$$

Gate j passes when $p_j \le \alpha_j$, for a prespecified error level α_j. In words, at $\alpha_j = 0.10$, the gate passes only if the new incident looks better than 90% of the past failures of the same type, because a failure that looks this good is rare. Conformal testing supplies the ranks (Vovk et al., 2005; Shafer & Vovk, 2008; Bates et al., 2023); causal compatibility defines membership in each null.

Assumption 1 (Null exchangeability and fixed pipeline)**.** *Within each null population, calibration and test incidents are exchangeable. The same prespecified proposal, retrieval, scoring, selection, querying, and tie-breaking rules are applied throughout calibration and testing.*

Calibration labels come from interventions, independent audit, SCMs, or finalized historical outcomes withheld from workflow evidence; deployment needs no label. The smallest rank, $1/(n_j+1)$, permits passing only if $n_j \ge \lceil 1/\alpha_j \rceil - 1$, without ensuring power. The gate statistics may also be learned on separate development data and frozen before calibration; the *learned-space* variant used in Section 5.4 replaces only the space statistic, and the guarantee is unchanged (Appendix B.4). Section 4 establishes error control.

3.3 Stages III–IV: resolve uncertainty and return a decision

If symptoms fit nothing on the list, reconsider the list; if flu and COVID both fit, seek evidence that distinguishes them before committing to a diagnosis. CAVEAT uses the two gates' results to decide what happens next. If the space gate fails, it withholds commitment and requests revision. If the gate passes, several candidates may still fit, so it retains the **uncertainty set**

$$S(O) := \{h \in \mathcal{H}(O) : \widehat{\Delta}_{\mathcal{Q}}(h; O) - A_V(O) \le \tau_{\text{set}}\}. \tag{5}$$

We calibrate τ_{set} from a designated compatible candidate's excess discrepancy on test-disjoint incidents. Under Appendix B.3's conditions, the set retains that candidate with marginal probability at least $1 - \alpha_{\text{set}}$. Commitment requires both gates to pass and $S(O) = \{\widehat{h}\}$; otherwise CAVEAT withholds commitment. The set supports further evidence acquisition or the optional action guard, not commitment to an unresolved cause. The optional action guard releases an action only if its estimated regret is at most ρ, the largest excess expected loss the decision maker tolerates, under every retained candidate (Appendix B.1).

Revision and evidence acquisition. Revision adds, replaces, or modifies candidates in $\mathcal{H}(O)$. To guide it, the **certificate** $C(O)$ reports the space-gate result, candidate-by-query discrepancies, closest-fitting candidates, and poorly explained queries (Appendix B.5). To distinguish unresolved candidates, CAVEAT chooses $q^+(O) \in \arg\max_{q \in \mathcal{Q}^+(O)} \{U(q; S(O), O) - \lambda c(q)\}$, where $\mathcal{Q}^+(O)$ contains feasible queries, U is prespecified expected separation utility, c is acquisition cost, and $\lambda \ge 0$ sets the trade-off; in the hypothetical visual-loss case, an MRI is useful if it distinguishes the remaining candidates enough to justify its cost. Revision changes the space; queries augment O. Both updates return to verification (Algorithm 1, Appendix B.1; practical requirements: Appendix B.6).

4 VERIFICATION, EVIDENCE ACQUISITION, AND DECISION CONSEQUENCES

Our theory concentrates on the guarantees provided by CAVEAT, which include the largest fraction of wrong answers the gates can pass (Theorem 1), the conditions under which that control still holds (Lemma 1, Proposition 3), the largest fraction of right answers any rule could retain (Proposition 1), and when the space verdict improves recovery over a fixed repair (Proposition 2). In addition, we further decompose the decision regret of a wrong candidate into the two failure types (Proposition 4), so that CAVEAT can still release an action when it is safe under every retained candidate.

Stage II: finite-sample dual-null safety. Theorem 1 bounds how often each gate lets its own failure type through, for any calibration size (Vovk et al., 2005; Bates et al., 2023).

Theorem 1 (Finite-sample dual-null control)**.** *Under Assumption 1, the ranks in Equation* (4) *satisfy, for each* $j \in \{V, S\}$,

$$\Pr\Big(p_j(O) \le \alpha_j \mid H_0^j\Big) \le \beta_j := \frac{\lfloor \alpha_j(n_j+1) \rfloor}{n_j+1}. \tag{6}$$

The floor counts the ranks that can pass: $\beta_j \le \alpha_j$, with no passes when calibration is too small. Every incompatible selection belongs to exactly one failure type, $H_0 := \{\widehat{h} \notin \mathcal{A}_\epsilon(O, \mathcal{Q})\} = H_0^V \,\dot\cup\, H_0^S$, where $\dot\cup$ denotes disjoint union. Let $\pi_j := \Pr(H_0^j \mid H_0)$ be each type's proportion among incompatible selections, so $\pi_V + \pi_S = 1$. The gatewise bounds give

$$\Pr(\mathsf{Commit} \mid H_0) \le \pi_V \beta_V + \pi_S \beta_S \le \max\{\beta_V, \beta_S\}. \tag{7}$$

The probability of a commitment conditional on an incompatible selection is capped at α when both levels are at most α.

Calibration follows the pipeline and its null. A stopping rule is needed for candidate expansion or revision, since unlimited revision raises the risk of false acceptance under reused calibration, and calibration must be consistent with the same revision and stopping rules. Under Lemma 1's conditions, reusing the threshold bounds the failure pass rate by $\min\{1, \beta_n + d_+\}$. Here β_n is the attainable level from the n source calibration incidents, and d_+ is the largest increase, over score cutoffs, in the probability that a failure passes once the procedure changes. If the distribution of failure scores is stable, then $d_+ = 0$. If the change makes failures more aligned with evidence, the bound grows (Appendix A.3; Barber et al., 2023). Section 5.4 tests this effect under adaptive search. Calibration must also match the failure type. A cutoff borrowed from the other failure type can withhold more compatible selections or increase false acceptance enough to exceed the level (Proposition 3, Appendix A.2).

Stages I–III: proposal and evidence jointly determine power. Besides controlling wrong commitment, we also care about how often compatible selections can be accepted. This requires the proposer to supply one supported by the evidence. Let γ_{a} be the probability of an adequate space, γ_{s} be the probability of a compatible selection given adequacy, and $\gamma_{\mathrm{prop}} = \gamma_{\mathrm{a}}\gamma_{\mathrm{s}}$ be their product. Let $F(O)$ be the evidence available at commitment, with distributions P_1^F and P_0^F under compatible and incompatible selection (Appendix A.4).

Proposition 1 (Proposal–verification operating region)**.** *Let* β *be the maximum permitted commitment probability conditional on an incompatible selection, and* Φ_β *the set of possibly randomized commitment rules* $\varphi(F) \in [0,1]$ *with* $\mathbb{E}_{P_0^F}[\varphi(F)] \le \beta$. *Let* $\mathcal{C}_F(\beta)$ *be the largest retention among them, attained by a Neyman–Pearson likelihood-ratio rule. Then*

$$\begin{aligned} \mathcal{C}_F(\beta) := \sup_{\varphi \in \Phi_\beta} \mathbb{E}_{P_1^F}[\varphi(F)] \;\le\; \min\big\{1, \beta + \mathrm{TV}(P_1^F, P_0^F)\big\}, \\ \sup_{\varphi \in \Phi_\beta} \Pr\Big(\widehat{h} \in \mathcal{A}_\epsilon(O, \mathcal{Q}), \mathsf{Commit}_\varphi\Big) = \gamma_{\mathrm{prop}}\, \mathcal{C}_F(\beta). \end{aligned} \tag{8}$$

Meaning. The proposer selects a compatible candidate with probability γ_{prop}; a verifier retains at most a fraction $\mathcal{C}_F(\beta)$ of these while accepting at most a β fraction of incompatible selections. Thus its false-refusal rate is at least $1 - \mathcal{C}_F(\beta)$. Total variation TV measures how differently the evidence looks under compatible and incompatible selections. If it looks the same in both cases,

no rule can retain more than a β fraction, however it is built. The likelihood-ratio optimum is classical (Neyman & Pearson, 1933). For unresolved candidates, useful acquisition must reveal a difference the current evidence cannot, and a follow-up query can improve localization only if its outcome has different laws under the candidates it is meant to separate (Corollary 1, Appendix A). For a fixed task, incident distribution, and selection, outcome W gives $F^+ = (F, W)$ and ideal gain $\gamma_{\mathrm{prop}}[\mathcal{C}_{F^+}(\beta) - \mathcal{C}_F(\beta)] \geq 0$ at the same error budget (Lemma 2; Blackwell, 1953). The gain is zero when, given the existing evidence, the query's outcome has the same law under compatible and incompatible selections.

Stage III: the space verdict guides repair and recovery. The *failure-matched* repair revises an inadequate space and queries an adequate one. We compare one-repair policies with shared early stopping and identical downstream rules for each branch, including terminal verification. Appendix A.9 extends the comparison to policy-specific terminal gates, as used in the trial. Let β_V be the routing gate's attainable level, $w_V := \Pr(H_0^V)$, $w_A := 1 - w_V$, and $\phi_V := \Pr(p_V > \alpha_V \mid (H_0^V)^c)$, the space gate's false-rejection rate on adequate initial spaces. Let $Y_{\mathrm{R}}, Y_{\mathrm{Q}}, Y_{\mathrm{route}} \in \{0, 1\}$ indicate correct commitment under revision, querying, and gate-based routing, respectively. Define the repair advantages $\Delta_V := \mathbb{E}[Y_{\mathrm{R}} - Y_{\mathrm{Q}} \mid H_0^V]$ and $\Delta_A := \mathbb{E}[Y_{\mathrm{Q}} - Y_{\mathrm{R}} \mid (H_0^V)^c]$.

Proposition 2 (Repair identification and routing value)**.** *Under Assumption 1, suppose querying adds evidence without changing the candidate space, task, or which represented candidates are task-compatible. Then (i) under H_0^V, querying cannot yield a correct commitment; (ii) an inadequate space escapes revision with probability at most β_V; (iii) the space verdict disagrees with the initial adequacy status with probability at most $w_V\beta_V + w_A\phi_V$; and (iv) under the shared branch rules above,*

$$\begin{aligned} &\mathbb{E}[Y_{\mathrm{route}}] - \max\{\mathbb{E}[Y_{\mathrm{R}}], \mathbb{E}[Y_{\mathrm{Q}}]\} \\ &\qquad \geq \min\{w_V\Delta_V, w_A\Delta_A\} - w_V\beta_V - w_A\phi_V. \end{aligned} \tag{9}$$

Routing therefore beats the best fixed repair when the benefit of matching repairs to space adequacy exceeds the bound on routing mistakes. Calibration controls the inadequate-space error, not the adequate-space false-rejection rate ϕ_V or repair effectiveness. Appendix A.9 gives the exact decomposition and an extension for policy-specific terminal gates. The gold-chosen reference in Section 5.3 additionally uses realized repair outcomes and is not the failure-matched repair.

Stage IV: hypothesis-space and selection failures create distinct harm. A wrong explanation matters because of the action it recommends. Decision regret is its excess expected loss over the operative mechanism's best action. It splits into two parts, $R_D = R_V + R_S$ (Proposition 4, Appendix A.8). R_V is the gap between the best action the space supports and the unrestricted best action; R_S is the additional loss from the selected action. Reranking can reduce R_S but cannot recover an action the space does not support. So the gate outcome decides whether to act. If the space is rejected, R_V may be large, so no action is taken. If the space passes but several candidates remain, R_S should be controlled, so the action guard releases an action only when its estimated regret is at most ρ under every retained candidate (Equation (20), Appendix A.10). A verified selection can support commitment; any released action must also satisfy the action guard.

5 Experiments: detection, repair, and the bounds in practice

We begin by testing our motivation that hypothesis-space inadequacy is common in real settings and costly when a proposer must commit. We then assess whether CAVEAT's two gates reduce those costs and its space gate picks the appropriate repair. Afterwards, we examine the guarantees of Section 4 against the observed rates. Finally, we report where CAVEAT stops helping and what limits it.

5.1 Settings, protocol, and metrics

Table 1 lists the settings, which span three kinds of evidence. *Simulated mechanisms* (SCMs and the CLadder extension) supply exact failure labels and unlimited probes. An SCM incident asks which mechanism produced the observed intervention responses, and a CLadder world asks which causal graph explains a story and its probed effects, with open 3B–32B models as proposers (Jin

et al., 2023; Yang et al., 2024). This completeness allows us to test CAVEAT's guarantees and the repair choice. *Real measurements under known interventions* (light and wind tunnels; Jurkat and HepG2 CRISPRi screens) ask which randomized intervention family produced a sensor or gene-expression response (Gamella et al., 2025; Nadig et al., 2025), helping us explore how CAVEAT improves recovery as more evidence is queried. *Real clinical records* (LiveClin) provide published patient cases with recorded diagnoses (Wang et al., 2026), and show us that CAVEAT is constrained by the quality of the evidence and the scorer. Inadequacy there is natural, since the retriever that proposes the space often misses the diagnosis. Except on LiveClin, inadequacy is constructed before testing. Controlled omissions withhold the true mechanism or family from the space while keeping its measured response. For physical, biological, and clinical benchmarks, compatibility follows the generating-family or recorded-diagnosis label (Appendix C). In the SCMs we further separate *detectable* omissions, which fit no observed consequence, from *indistinguishable* ones, which differ from the remaining candidates only on an unmeasured task query.

Table 1: **Evaluation settings.** Proposers and scorers are frozen before incident- or family-disjoint calibration and testing; splits, omission constructions, and compatibility definitions are in Appendix C.

Setting	Test units	Proposer	Frozen scorer
Controlled SCMs	8,000 incidents per omission type	fit ranking over the library	squared response error
CLadder extension	120 worlds; matched trial 240	Qwen2.5-7B (3B–32B ladder)	squared residual of predicted effect
Light / wind tunnel	29 / 21 families; 203 / 105 omitted roles	fit ranking; fingerprint retrieval	standardized sensor error
Jurkat / HepG2 screens	720 / 360 units; 180 families each	fit ranking over families	multiscale profile discrepancy
LiveClin	171 test, 167 calibration cases	MedCPT cross-encoder, top 10	negative cross-encoder support

Baselines and ablations. We compare each baseline proposer with the same proposer plus CAVEAT, using the fixed scorer, on two functions: reducing wrong commitment while keeping the correct answers that pass both gates, and how the two repairs, revise and query, change the proposer's performance. Ablations vary the discrepancy score and mask one component at a time, such as the selection gate or the learned score (Appendix D.7).

Metrics. Wrong and Correct are the fractions of all test units, noncommitments included, committed to incompatible and compatible candidates, respectively; "under omission" restricts the units to inadequate spaces. For LLM repair workflows, retention and recovery are the fractions of initially compatible and incompatible selections, respectively, that end in correct commitment. Chamber recovery instead measures correct retrieval among omitted roles, before commitment. Verification-only retention conditions on the fixed selection (Appendix C).

5.2 Detection: inadequacy is costly, and verification limits wrong commitment

Setting	Units	Omission units	Forced choice	CAVEAT, both gates
Light tunnel	638	203 (32%)	71.2% [67.1, 75.2]	0.5% [0.0, 1.4]
Wind tunnel	336	105 (31%)	49.1% [46.1, 52.1]	0.3% [0.0, 0.9]
Jurkat	720	360 (50%)	83.9% [80.7, 86.8]	1.1% [0.3, 2.2]
HepG2	360	180 (50%)	75.3% [71.7, 78.9]	1.7% [0.3, 3.6]
LiveClin	171	133 (78%)	93.6% [88.8, 96.7]	0.0% [0.0, 2.1]
CLadder	120	60 (50%)	75.8% [67.2, 83.2]	0.8% [0.0, 4.6]
Matched trial, 7B	240	122 (51%)	76.2% [70.5, 81.2]	1.7% [0.6, 4.2]
Matched trial, 14B	240	122 (51%)	67.9% [61.8, 73.5]	2.1% [0.9, 4.8]
SCM, detectable	8,000	8,000 (100%)	100.0%	0.0% [0.0, 0.05]
SCM, indistinguishable	8,000	8,000 (100%)	100.0%	0.48% [0.34, 0.65]

Table 2: **Wrong commitment under forced choice and verified workflows.** Omission units contain no task-compatible candidate, so any commitment is wrong. Rates use all test units, with 95% intervals reported in the source archives; the matched trial uses Wilson intervals. Chamber compares three-sensor forced choice with all-sensor verification; LLM workflows also include repair. Sources: Appendix D.7 and Table 16.

Finding 1: forced choice pays for inadequacy; verification limits wrong commitment. Table 2 compares wrong commitment under forced choice and the verified workflow in each setting. The omission proportion is high, from about a third in the chambers to 78% on LiveClin, and forced choice is wrong on every such unit, so its wrong commitment reaches 49–94% of all units on the real-data settings. Omission units account for 45–83% of forced choice's errors in these settings. With CAVEAT, wrong commitment falls to 0–2% in these primary comparisons. Withholding on a single score is not enough: on LiveClin, margin, entropy, and one-score gates still commit wrongly in 8–28

of 171 cases, whereas CAVEAT and ordinary conformal commit wrongly in none (Table 4A). In the SCMs, no detectable omission passes the space gate. Among the indistinguishable omissions 20.2% pass it, but only 0.48% are wrongly committed because the selection gate withholds the rest. The test panel mixes both kinds equally, giving an observed space-gate pass rate of 10.1% at nominal 10%.

Model	CAVEAT	Revise	Query	Random
Qwen 3B	79/1	53/2	73/3	61/2
Qwen 7B	100/1	57/1	90/5	70/5
Qwen 14B	123/2	77/2	109/4	95/3
Qwen 32B	131/5	84/4	111/5	100/5
Phi-3.5	86/2	61/2	75/5	69/4
OLMo-2	51/3	44/3	42/6	50/5

Table 3: **Routing when revision can fail.** Correct/wrong commitment counts out of 240 worlds at .50 compatible-candidate supply probability. Revise and Query always choose that repair; Random chooses between them. Policies share one repair opportunity and receive matched terminal calibration at .10/.10. All models share the worlds; inference settings are in Appendix C.2, and full supply curves are in Figure 3.

5.3 Repair: acting on the detection recovers answers where the evidence allows

Finding 2: the space gate guides the repair, and both gates verify the result. In the controlled 120-world CLadder experiment, the 7B model's self-reflection changes neither retained answers nor wrong commitments (Figure 2A). Raw errors remain frequent across 3B–32B models (Appendix C.1). Revision instead supplies missing candidates, and queried evidence helps resolve alternatives. CAVEAT reaches 42 correct commitments, comprising 25 retained and 17 recovered answers, with one wrong against 91 wrong under raw forced choice (Appendices D.7, D.8).

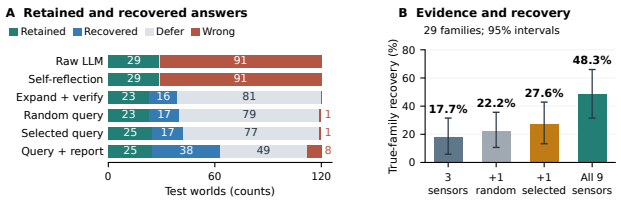

Figure 2: **CLadder feedback and Chamber recovery.** A (CLadder): 120 worlds. Retained: the proposer's original correct choice is kept; recovered: a wrong choice is corrected by the repair. Query arms use 0.81 extra probes on average. B (Chamber): light tunnel, 203 omitted roles in 29 families; omitted families recovered with three sensors, one more sensor chosen at random or by CAVEAT, or all nine.

We next examine the routing benefit in Proposition 2, comparing routing with always revising, always querying, and random repair on 240 new worlds. The trial controls candidate supply and separating probes; policies share initial evidence, one repair opportunity, and at most one extra model call. Routing yields 20 and 30 more correct commitments than the stronger fixed repair at 7B and 14B, and stays five and seven below the gold-chosen reference. Wrong commitments match the fixed repairs; versus random repair, routing adds one at 7B and none at 14B (Tables 11–12). The space-verdict audit is reported in Appendix C.1.

The repair need not always succeed for routing to help. On a fresh 240-world panel, we vary the probability that revision supplies a compatible candidate and recalibrate each policy for each condition. At 50% supply probability, routing adds 6–20 correct commitments over the stronger fixed repair across Qwen 3B–32B (Table 3). For 7B, its gain over random repair is 12.5 percentage points [7.9, 17.5] (95% paired story-cluster interval). Phi also improves over random repair, while OLMo is nearly tied (51 versus 50 correct). To examine whether the reasoners use the supplied evidence, a post-hoc audit at full supply selects the candidate with the uniquely lowest discrepancy. This candidate is task-compatible in 178 of 240 routed worlds for every reasoner, whereas the LLMs select compatible candidates in only 54–147. Recovery therefore also depends on the reasoner's ability to select the evidence-supported candidate (Appendix C.2).

Finding 3: recovery follows the separation of the evidence, and where it is weak CAVEAT withholds. Additional sensors improve light-tunnel recovery (Figure 2B), but the sealed wind-tunnel study fails its prespecified recovery and query-utility criteria (Appendix D.5). LiveClin shows a different limit. With the tested scorer and available records, CAVEAT withholds almost every exact commitment (Table 4A) and instead returns uncertainty sets or broad directions (Table 4B–C). The directions are hierarchical conformal sets, so on LiveClin the main thing CAVEAT adds is the

Table 4: **LiveClin: 171 sealed test cases.** A: exact decisions per policy, counts out of 171 with exact 95% binomial intervals; None means no exact answer was committed. B: CAVEAT's output, an exact answer, an uncertainty set, or a broad direction (a set of level-2 diagnostic categories), split by whether the recorded diagnosis was in the retrieved space. C: coverage of the recorded category and mean size of the directions, which CAVEAT builds with the standard hierarchical conformal construction over the 72 level-2 categories. CAVEAT requests 2.82 further evidence items per case on average.

A. Exact decisions, count/171 [95% interval, %]

Policy	Correct ↑	Wrong ↓	None
Forced, initial evidence	11/171 [3.3, 11.2]	160/171 [88.8, 96.7]	0
Forced, full evidence	12/171 [3.7, 11.9]	159/171 [88.1, 96.3]	0
Margin abstention	4/171 [0.6, 5.9]	28/171 [11.2, 22.8]	139
Entropy abstention	5/171 [1.0, 6.7]	28/171 [11.2, 22.8]	138
One-score gate	1/171 [0.0, 3.2]	8/171 [2.0, 9.0]	162
Ordinary conformal	2/171 [0.1, 4.2]	0/171 [0.0, 2.1]	169
CAVEAT	1/171 [0.0, 3.2]	0/171 [0.0, 2.1]	170

B. CAVEAT output by space state

Recorded diagnosis	Exact	Set	Direction
Absent (133 cases)	0	10	123
Present (38 cases)	1	1	36

C. Directions, 159 cases, 72 categories

Policy	Coverage [95%]	Size ↓
CAVEAT	146/159 [86.4, 95.6]	32.7

refusal to commit, not the set itself. Even when the missing diagnosis is added, the scorer still produces no exact singleton (Appendix D.6). Adding the right candidate does not resolve indistinguishable alternatives (Proposition 1). Figure 2B also shows no significant improvement from CAVEAT's selected query over a random query when several useful sensors are available (Table 17A), possibly because the tunnel sensors provide similar amounts of information.

5.4 Bounds: the theory's inequalities against the observed rates

Finding 4: calibration, the level α, and the gate statistics decide how many correct answers are kept. In the exposed K562 CRISPRi development sweep, correct commitment rises from 11.6% to 30.8% as α rises from .05 to .20, while wrong commitment under omission rises from 2.6% to 17.2%. CAVEAT exceeds the matched baselines at each tested level (Figure 4). A development-trained space score also improves retention at fixed levels in fresh simulation (Table 7). These gains require matched calibration: reusing a pre-search cutoff raises wrong commitment from 10.4% to 43.7% in the adaptive SCM. Table 13 in Appendix D compares empirical rates with the corresponding theoretical bounds, using the conditioning specified for each guarantee.

Finding 5: the action guard adds a second safety layer, on the harm of acting. Beyond the two gates on the cause, CAVEAT checks the action itself (Proposition 4) against a predefined regret tolerance ρ. When several candidates remain in the uncertainty set, it releases an action only if that action is within tolerance under every retained candidate (Equation (20), Appendix A.10). Among 86,400 factorial SCM incidents, forced action without verification exceeds the tolerance 10,780 times (12.5%). CAVEAT releases 1,332 actions with none exceeding the tolerance, and 610 of them are released with two candidates still retained. CAVEAT performs more cautiously by withholding actions that are in fact safe whenever they fall outside its prespecified level α or tolerance ρ; raising either releases more of them under a correspondingly looser guarantee (Table 16).

6 Impact, limitations, and conclusion

We formalize hypothesis-space inadequacy and design CAVEAT around two calibrated gates that distinguish it from selection failure and guide their respective repairs. Our theory bounds acceptance of incompatible selections and gives a sufficient condition for test-guided repair to outperform either fixed repair. Across the tested settings, verification reduces wrong commitment, while retained usefulness depends on the evidence and scorer. In controlled LLM reasoning, routing improves recovery under imperfect candidate supply for Qwen and Phi, while OLMo remains near random repair. The selection audit shows that recovery also requires the reasoner to choose the evidence-supported candidate. Limitations include few independent physical families, retrospective clinical evaluation, and learned scores tested only in simulation (Appendix D.9). The LLM trials use controlled candidate supply and query evidence, and the action guard is worst-case. The guarantees are population-level and require recalibration under distribution shift (Barber et al., 2023). These limitations motivate open causal benchmarks with interventional ground truth to distinguish what better scoring can resolve from what requires new evidence. The broader lesson is that diagnostic systems should test their hypothesis space before trusting its ranking.

AI USE STATEMENT

Generative AI tools (Anthropic Claude and OpenAI Codex) assisted with writing polish, prototyping code for experimental design, and debugging. In addition, We also used generative AI tools as critical readers to help check and simplify our theorem proofs. The authors made all final research decisions, checked citations, and reported values against primary sources and frozen archives, reviewed all AI-assisted text, code, and mathematics, and take responsibility for the final content.

ETHICS STATEMENT

LiveClin consists of published case reports released under a CC-BY-NC-SA license; no patient contact or clinical deployment is involved, and the study is a retrospective benchmark stress test, not clinical validation or medical advice. The CRISPRi screens and Causal Chamber data are public. CAVEAT withholds rather than issues causal commitments when its tests fail; a deployment that removed the calibration or exchangeability conditions of Section 3.2 would forfeit the guarantee.

REPRODUCIBILITY STATEMENT

The supplementary material contains the reference Python package, frozen result archives, and experiment and reproduction scripts. The package provides full-workflow calibration, verification, structured reports, and tested integration examples. Its `01_reference_package/README.md` describes the input contract, calibration requirements, and execution instructions. The function `build_calibration_set` validates labelled incidents and constructs $\mathcal{C}_V$ and $\mathcal{C}_S$. Frozen-result checks require no new model calls; the cluster notebooks document the original inference procedures. Any change to the candidate generator, prompt, retrieval depth, discrepancy score, query rule, or tie-breaking convention defines a new pipeline and requires fresh calibration.

REFERENCES

Ahmed Abdulaal, Adamos Hadjivasiliou, Nina Montaña-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fe90657b12193c7b52a3418bdc351807-Abstract-Conference.html.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://doi.org/10.1561/2200000101.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276. URL https://arxiv.org/abs/2202.13415.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023. doi: 10.1214/22-AOS2244. URL https://doi.org/10.1214/22-AOS2244.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1563–1572, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Bendale_Towards_Open_Set_CVPR_2016_paper.html.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406. URL https://doi.org/10.1109/TIT.1970.1054406.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020. doi: 10.1111/rssb.12348.

Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/bc75254bc4b8b42f401d0ab5d6e9aa4b-Abstract-Conference.html.

Philipp M. Faller, Leena C. Vankadara, Atalanti A. Mastakouri, Francesco Locatello, and Dominik Janzing. Self-compatibility: Evaluating causal discovery without ground truth. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4132–4140. PMLR, 2024. URL https://proceedings.mlr.press/v238/faller24a.html.

Aoyang Fang, Yifan Yang, Jin'ao Shang, Qisheng Lu, Junjielung Xu, Rui Wang, Songhan Zhang, Yuzhong Zhang, Boxi Yu, and Pinjia He. OpenRCA 2.0: From outcome labels to causal process supervision. *arXiv preprint arXiv:2606.27154*, 2026. URL https://arxiv.org/abs/2606.27154.

Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7:107–118, 2025. doi: 10.1038/s42256-024-00964-x. URL https://doi.org/10.1038/s42256-024-00964-x.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–807, 1996. URL https://www3.stat.sinica.edu.tw/statistica/j6n4/j6n41/j6n41.htm.

Mark L. Graber, Nancy Franklin, and Ruthanna Gordon. Diagnostic error in internal medicine. *Archives of Internal Medicine*, 165(13):1493–1499, 2005. doi: 10.1001/archinte.165.13.1493. URL https://pubmed.ncbi.nlm.nih.gov/16009864/.

Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014. doi: 10.1016/j.ijar.2013.11.007. URL https://doi.org/10.1016/j.ijar.2013.11.007.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/631bb9434d718ea309af82566347d607-Abstract-Conference.html.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Xiuchuan Li, Kun Zhang, and Tongliang Liu. Causal structure recovery with latent variables under milder distributional and graphical assumptions. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/730ce0ae730f39e4d77b0f04a8afe4be-Abstract-Conference.html.

Ye Li, Anqi Hu, Yuanchang Ye, Shiyan Tong, Zhiyuan Wang, and Bo Fu. Set-valued prediction for large language models with feasibility-aware coverage guarantees. *arXiv preprint arXiv:2603.22966*, 2026. URL https://arxiv.org/abs/2603.22966.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. Discovery of the hidden world with large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3249. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b99a07486702417d3b1bd64ec2cf74ad-Abstract-Conference.html.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

MedlinePlus Medical Encyclopedia. Optic neuritis. National Library of Medicine, 2026. URL https://medlineplus.gov/ency/article/000741.htm. Reviewed January 27, 2026. Accessed September 7, 2026.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36029–36047. PMLR, 2024. URL https://proceedings.mlr.press/v235/mohri24a.html.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Ajay Nadig, Joseph M. Replogle, Angela N. Pogson, Mukundh Murthy, Steven A. McCarroll, Jonathan S. Weissman, Elise B. Robinson, and Luke J. O'Connor. Transcriptome-wide analysis of differential expression in perturbation atlases. *Nature Genetics*, 57(5):1228–1237, 2025. doi: 10.1038/s41588-025-02169-3. URL https://www.nature.com/articles/s41588-025-02169-3.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231(694–706): 289–337, 1933. doi: 10.1098/rsta.1933.0009. URL https://doi.org/10.1098/rsta.1933.0009.

William Roy Orchard, Philipp Michael Faller, and Dominik Janzing. Falsifying causal graphs with outlier events. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 5081–5110. PMLR, 2026. URL https://proceedings.mlr.press/v337/orchard26a.html.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. ISBN 978-0-521-89560-6.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008. URL https://www.jmlr.org/papers/v9/shafer08a.html.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2 edition, 2000. ISBN 978-0-262-19440-2.

Zehao Su and Leonard Henckel. A robustness test for estimating total effects with covariate adjustment. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1886–1895. PMLR, 2022. URL https://proceedings.mlr.press/v180/su22a.html.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840. PMLR, 2022. URL https://proceedings.mlr.press/v162/sun22d.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, NY, 2005. ISBN 978-0-387-25061-8. doi: 10.1007/b106715. URL https://doi.org/10.1007/b106715.

Xidong Wang, Shuqi Guo, Yue Shen, Junying Chen, Jian Wang, Jinjie Gu, Ping Zhang, Lei Liu, and Benyou Wang. LiveClin: A live clinical benchmark without leakage. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9f45b2d4ace7ae425e40c7c1d0d37e64-Abstract-Conference.html.

Tianmin Xie, Yanfei Zhou, Ziyi Liang, Stefano Favaro, and Matteo Sesia. Conformal inference for open-set and imbalanced classification. *arXiv preprint arXiv:2510.13037*, 2025. URL https://arxiv.org/abs/2510.13037.

Jian Xu, Yanning Wu, Delu Zeng, John Paisley, and Qibin Zhao. Look again before you abstain: Budgeted conformal evidence acquisition for reliable vision-language model. *arXiv preprint arXiv:2606.16667*, 2026. URL https://arxiv.org/abs/2606.16667.

Junjielong Xu, Qinan Zhang, Zhiqing Zhong, Shilin He, Chaoyun Zhang, Qingwei Lin, Dan Pei, Pinjia He, Dongmei Zhang, and Qi Zhang. OpenRCA: Can large language models locate the root cause of software failures? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d29b8d53678015079e1d245c023e49d2-Abstract-Conference.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

A Proofs

A.1 Proof of Theorem 1

Fix $j \in \{V, S\}$ and suppose the test incident is a failure of type j. Under Assumption 1, its score and the n_j calibration scores are exchangeable. Assign independent $\mathrm{Uniform}(0,1)$ auxiliary variables to all scores, independently of the scores, and use them to break ties. The resulting test rank R is uniform on $\{1, \dots, n_j + 1\}$. The weak rank used by CAVEAT places the test score at the end of its tie block, so $p_j(O) \geq R/(n_j + 1)$. Therefore,

$$\Pr(p_j(O) \leq \alpha_j \mid H_0^j) \leq \Pr\Big(R \leq \lfloor \alpha_j(n_j+1) \rfloor \mid H_0^j\Big) = \frac{\lfloor \alpha_j(n_j+1) \rfloor}{n_j+1} = \beta_j.$$

The auxiliary variables are used only in the proof; the implemented weak rank remains deterministic and conservative under ties.

Composite consequence. Let E be the commitment event. By Algorithm 1 (Appendix B.1), commitment requires both gates to pass, so $E \subseteq \{p_V \le \alpha_V\}$ under H_0^V and $E \subseteq \{p_S \le \alpha_S\}$ under H_0^S. The bound above then gives $\Pr(E \mid H_0^V) \le \beta_V$ and $\Pr(E \mid H_0^S) \le \beta_S$. Since every incompatible selection belongs to exactly one of H_0^V and H_0^S, the law of total probability gives Equation (7), and multiplying by $\Pr(H_0)$ gives $\Pr(\mathsf{Commit}, H_0) \le \Pr(H_0) \cdot \max\{\beta_V, \beta_S\}$.

A.2 Null-matched calibration (Proposition 3)

The selection gate needs its own bank of past selection failures, which are scarcer than space failures. One might be tempted to calibrate the selection threshold on the space-failure bank instead, so that both gates share one population. Proposition 3 shows that this shortcut carries no guarantee. Depending on where the borrowed threshold falls, it either refuses compatible selections it should accept or lets the level be exceeded, and Appendix D.7 observes both outcomes on the chambers. This is why CAVEAT calibrates each gate on its own failures.

We now state the proposition and its proof. Write $M_S(O) := -A_S(O)$ for the selection margin oriented so that larger values support the selected candidate; it is positive when the selection fits better than its best alternative and negative when it fits worse. The gate passes when M_S exceeds a threshold. Call t_m the threshold calibrated on selection failures and t_b the one borrowed from space failures.

Proposition 3 (Null-matched calibration)**.** *Let P_0^S be the law of M_S under H_0^S, and P_1 its law under compatible selection, and write $r_m = P_0^S(M_S > t_m)$ for the risk of the matched threshold. If $t_b \ge t_m$, borrowing cannot increase risk and decreases power by $P_1(t_m < M_S \le t_b)$. If $t_b < t_m$, its risk is $r_m + P_0^S(t_b < M_S \le t_m)$, which exceeds α exactly when the added mass exceeds $\alpha - r_m$.*

Proof. The gate accepts on $\{M_S > t\}$. If $t_b \ge t_m$, the borrowed acceptance event is a subset of the matched one and their difference is the strip $\{t_m < M_S \le t_b\}$. If $t_b < t_m$, the borrowed event is the disjoint union of the matched one and the strip $\{t_b < M_S \le t_m\}$. Taking probabilities under P_1 and P_0^S gives the two statements. For data-dependent thresholds the identities hold conditional on the calibration data and can be averaged. Theorem 1 is a marginal bound, so ties or conservative calibration do not justify replacing r_m by α.

A.3 Calibration transfer for a frozen gate

A threshold is calibrated on the failure scores of one pipeline. Revision, adaptive search, or any other change to the pipeline changes the law of those scores, and recalibrating after every change is costly. Thus, the question is what happens if we reuse a threshold without any adjustment to the change. Lemma 1 answers it for either gate under its own null. Unlike Theorem 1, the lemma uses independent calibration and test sampling, with training and design choices held fixed.

We now state the lemma and its proof, and first fix the notation. The *source* is the pipeline that supplied the calibration scores and the *target* is the changed pipeline. Their failure scores and curves are

$$\text{source: } \{A_1, \dots, A_n\} \;\to\; F_{\mathcal{P}}(t) = \Pr_{\mathcal{P}}(A_j < t \mid H_0^j),$$
$$\text{target: } \{A'_1, A'_2, \dots\} \;\to\; F_{\mathcal{P}'}(t) = \Pr_{\mathcal{P}'}(A_j < t \mid H_0^j),$$

where $F(t)$ is the fraction of failures scoring below t. The gate keeps the source threshold $T = A_{(k)}$, the kth smallest source score, and passes a target score when it falls strictly below T. Calibration gives $\mathbb{E}[F_{\mathcal{P}}(T)] \le \beta_n$; the target pass rate at threshold T is $F_{\mathcal{P}'}(T)$; and

$$d_+ = \sup_t [F_{\mathcal{P}'}(t) - F_{\mathcal{P}}(t)]_+$$

is the largest amount by which the target curve rises above the source curve at any t, so the target pass rate exceeds the source pass rate by at most d_+.

Lemma 1 (Finite-sample calibration transfer)**.** *Fix gate $j \in \{V, S\}$ and frozen pipelines $\mathcal{P}, \mathcal{P}'$ with strict null distribution functions $F_{\mathcal{P}}(t) = \Pr_{\mathcal{P}}(A_j < t \mid H_0^j)$ and $F_{\mathcal{P}'}(t) = \Pr_{\mathcal{P}'}(A_j < t \mid H_0^j)$. Let $A_1, \dots, A_n$ be iid source-null calibration scores and A' an independent target-null score. For $\alpha \in (0,1)$ set $k = \lfloor \alpha(n+1) \rfloor$, $\beta_n = k/(n+1)$, and $T = A_{(k)}$ if $k \ge 1$, otherwise $T = -\infty$;*

the transferred strict gate passes when $A' < T$. With $[x]_+ = \max(x, 0)$ and $d_+ = \sup_t [F_{\mathcal{P}'}(t) - F_{\mathcal{P}}(t)]_+$,

$$\Pr(A' < T) = \mathbb{E}F_{\mathcal{P}'}(T) = \mathbb{E}F_{\mathcal{P}}(T) + \mathbb{E}[F_{\mathcal{P}'}(T) - F_{\mathcal{P}}(T)] \le \min\{1, \beta_n + d_+\}. \tag{10}$$

In particular $F_{\mathcal{P}'} \le F_{\mathcal{P}}$ preserves the source attainable level, and for a continuous source law $\mathbb{E}F_{\mathcal{P}}(T) = \beta_n$.

Proof. The target score is independent of the calibration scores, so conditioning on the threshold T gives $\Pr(A' < T) = \mathbb{E}F_{\mathcal{P}'}(T)$, the first equality. Adding and subtracting $F_{\mathcal{P}}(T)$ gives the second. For the first term, take an independent source-null score A_0; its strict rank among the $n + 1$ exchangeable source scores gives $\mathbb{E}F_{\mathcal{P}}(T) = \Pr(A_0 < T) \le k/(n + 1) = \beta_n$, with ties conservative and equality for a continuous law. For the second term, the difference of the two laws at any threshold is at most d_+ by definition. The bound is also at most one, and the $k = 0$ gate never passes.

The expectation is over calibration and test, so a realized threshold need not have source pass probability at most α. The directional discrepancy d_+ quantifies transfer of this fixed score and must itself be justified or estimated before it supports deployment. The result is a direct score-law comparison, complementary to conformal robustness under nonexchangeability (Barber et al., 2023), and for CAVEAT it separates the two null score laws that revision or search can change. Stochastic dominance is equivalent to non-increase of risk at every deterministic threshold and is not necessary for one conservative threshold to stay safe. For example, with $F_{\mathcal{P}}(t) = t$ and $F_{\mathcal{P}'}(t) = \sqrt{t}$ on $[0, 1]$, the threshold $t_\alpha = \alpha^2/4$ has target pass probability $\alpha/2 \le \alpha$ although $F_{\mathcal{P}'} > F_{\mathcal{P}}$ in the interior.

Search and the actual pass probability. Adaptive search is one way the score law changes. When a proposer searches over M candidates and keeps the best-fitting one, the score that reaches the gate is the minimum of M discrepancies, which passes more easily than any single one. For a fixed threshold t and M candidate discrepancies with source-null marginal law $F_{\mathcal{P}}$, writing $p = F_{\mathcal{P}}(t)$,

$$p \le \Pr\left(\min_{m \le M} A_m < t\right) \le \min(1, Mp), \tag{11}$$

by inclusion of each candidate event and the union bound; independence gives $1 - (1 - p)^M$ and perfectly dependent copies give p. For a random threshold independent of the candidates these hold conditional on the threshold and are then averaged. For M independent continuous source-law candidates and the kth-order threshold above,

$$\Pr\left(\min_{m \le M} A_m < T\right) = 1 - \prod_{j=0}^{M-1} \frac{n + 1 - k + j}{n + 1 + j}, \tag{12}$$

since $F_{\mathcal{P}}(T) \sim \mathrm{Beta}(k, n+1-k)$ for $k \ge 1$ and its Mth complementary moment gives the product (zero when $k = 0$). This generally differs from substituting α into the fixed-threshold formula. The adaptive SCM selects candidates with correlated proposal scores and family-shared evidence, so its inflation illustrates a changed null score law rather than this independent-candidate case.

A.4 Operating-region notation

Correct commitment needs a compatible candidate in the space, its selection, and evidence that distinguishes it from an incompatible one. Define

$$\gamma_{\mathrm{a}} := \Pr(\mathcal{A}_\epsilon(O, \mathcal{Q}) \ne \varnothing), \qquad \gamma_{\mathrm{s}} := \Pr\Big(\widehat{h} \in \mathcal{A}_\epsilon(O, \mathcal{Q}) \,\Big|\, \mathcal{A}_\epsilon(O, \mathcal{Q}) \ne \varnothing\Big), \qquad \gamma_{\mathrm{prop}} := \gamma_{\mathrm{a}}\gamma_{\mathrm{s}}, \tag{13}$$

and let $F(O)$ be everything available to the verifier at commitment (dual scores, uncertainty set, acquired evidence), with

$$P_1^F := \mathcal{L}\Big(F(O) \,\Big|\, \widehat{h} \in \mathcal{A}_\epsilon(O, \mathcal{Q})\Big), \qquad P_0^F := \mathcal{L}\Big(F(O) \,\Big|\, \widehat{h} \notin \mathcal{A}_\epsilon(O, \mathcal{Q})\Big), \tag{14}$$

both conditioning events of positive probability; a rule's expectations under P_0^F and P_1^F are its false-commitment rate and retention. The optimization fixes task, selection, and population and varies only the commitment rule.

A.5 Proof of Proposition 1

For any commitment rule φ measurable in the final verification statistic, $\Pr(\widehat{h} \in \mathcal{A}_\epsilon(O, \mathcal{Q}), \mathsf{Commit}_\varphi) = \gamma_{\mathrm{prop}} \mathbb{E}_{P_1^F}[\varphi(F)]$; maximizing subject to $\mathbb{E}_{P_0^F}[\varphi(F)] \le \beta$ gives $\gamma_{\mathrm{prop}} \mathcal{C}_F(\beta)$. For every rule $0 \le \varphi \le 1$,

$$\mathbb{E}_{P_1^F}[\varphi] \;\le\; \mathbb{E}_{P_0^F}[\varphi] + \mathrm{TV}(P_1^F, P_0^F) \;\le\; \beta + \mathrm{TV}(P_1^F, P_0^F),$$

where the first inequality is the variational form of total variation for functions bounded by one and the second is the constraint defining Φ_β; the left side is also at most one. The Neyman–Pearson lemma gives the likelihood-ratio-optimal rule, with boundary randomization when required.

A.6 Corollary 1 and its proof

Corollary 1 (Localization requires separating evidence)**.** *If $P_h^{\mathcal{Q}} = P_{h'}^{\mathcal{Q}}$, the joint laws of the available query evidence under h and h', no evidence-based rule can distinguish candidates h and h' better than chance under equal priors. A follow-up query q^+ can improve localization only if the conditional inequality below holds on a set of observations of positive probability:*

$$P_h^{\mathcal{Q}} = P_{h'}^{\mathcal{Q}}, \qquad P_{h,q^+}(\cdot \mid O) \neq P_{h',q^+}(\cdot \mid O). \tag{15}$$

Proof. If $P_h^{\mathcal{Q}} = P_{h'}^{\mathcal{Q}}$, the evidence has the same law under both candidates, so under equal priors any rule forced to choose between them succeeds with probability at most $1/2$. If the query outcome also has the same conditional law given the original evidence under both candidates, the joint laws remain equal; changing the information bound therefore requires different conditional laws on a set of positive probability, as in Equation (15).

A.7 Query value at a fixed commitment stage

A measurement is useful when it changes what the decision rule can retain at the required error level; we measure that ideal value with the Neyman–Pearson power function of Proposition 1, a binary comparison of experiments (Blackwell, 1953).

Lemma 2 (Budget-specific value of evidence)**.** *Fix the task, incident distribution, selected hypothesis, and compatibility event $C = \{\widehat{h} \in \mathcal{A}_\epsilon(O, \mathcal{Q})\}$ with $0 < \Pr(C) = \gamma_{\mathrm{prop}} < 1$. Let $F^+ = (F, W)$ retain the original information and add the outcome W of a prespecified query policy (W may include the queried variable and policy randomization; all information spaces standard Borel). With the laws of F, F^+ under C and C^c and power functions as in Equation* (8)*,*

$$\Delta_{\mathrm{query}}(\beta) := \gamma_{\mathrm{prop}} \left[\mathcal{C}_{F^+}(\beta) - \mathcal{C}_F(\beta)\right] \ge 0 \tag{16}$$

is the difference between optimal correct-commitment probabilities at false-commitment budget β. If a common conditional kernel $K(dw \mid f)$ generates W given $F = f$ under both C and C^c, then $\Delta_{\mathrm{query}}(\beta) = 0$ for every β; identical marginal laws of W alone are insufficient.

Proof. Every rule based on F is available with F^+ by ignoring W, so optimal power cannot decrease, and Proposition 1 gives Equation (16) because C and γ_{prop} are fixed. Under a common kernel, any randomized rule $\varphi^+(f, w)$ induces $\bar{\varphi}(f) = \int \varphi^+(f, w) K(dw \mid f)$ with the same expectation under each selection state, so every attainable risk–power pair after querying is attainable before it.

This is an information benchmark, not a guarantee that a learner or finite-calibration gate attains the frontier; query cost and the consumer's response determine realized utility, and if revision changes the selected hypothesis, task, or population, the compatibility event must be redefined for that stage. Two finite examples fix the meaning. Equal marginals can conceal a separating measurement: under C^c let (F, W) be $(0, 0)$ or $(1, 1)$ with equal probability and under C be $(0, 1)$ or $(1, 0)$; both marginals are fair bits in both states, so $\mathcal{C}_F(\beta) = \beta$, while the pair identifies the state, $\mathcal{C}_{F^+}(\beta) = 1$. Total variation is not a budget-specific value: for $P_0 = (1/2, 1/2)$, $P_1 = (3/4, 1/4)$ and a query refining the first outcome into two, $P_0^+ = (1/10, 4/10, 1/2)$, $P_1^+ = (3/10, 45/100, 1/4)$, both pairs have total variation $1/4$, yet at $\beta = 1/10$ likelihood-ratio ordering gives $\mathcal{C}_F(\beta) = 3/20$ and $\mathcal{C}_{F^+}(\beta) = 3/10$, so ideal correct commitment improves by $3\gamma_{\mathrm{prop}}/20$ at zero total-variation increment.

A.8 Action losses and regret decomposition

Let $\mathcal{U}$ be the action set, Y the outcome, and $P_h^u, P_\star^u$ its laws after action u under candidate h and the operative mechanism. For harm $\ell(u, Y)$, let $J_h(u) := \mathbb{E}_{Y\sim P_h^u}[\ell(u, Y)]$ (predicted harm), $J_\star(u) := \mathbb{E}_{Y\sim P_\star^u}[\ell(u, Y)]$ (population harm), and $u_h \in \arg\min_u J_h(u)$ (candidate-induced action), with $J^\star := \min_u J_\star(u)$, $J^\star_{\mathcal{H}} := \min_{h\in\mathcal{H}(O)} J_\star(u_h)$, and $J_{\mathrm{sel}} := J_\star(u_{\widehat{h}})$.

Proposition 4 (Hypothesis-space–selection regret decomposition)**.** *$R_V := J^\star_{\mathcal{H}} - J^\star$, $R_S := J_{\mathrm{sel}} - J^\star_{\mathcal{H}}$, and $R_D := J_{\mathrm{sel}} - J^\star = R_V + R_S$.*

Proof. Add and subtract $J^\star_{\mathcal{H}}$.

A.9 Proof of Proposition 2

All probabilities refer to the same initial incident population as the routing calibration, not a subgroup selected by a previous gate. Calibration randomness is included in the probabilities where relevant. The task and compatibility criterion are fixed; querying adds evidence but leaves the represented candidates and their task compatibility unchanged. Let

$$\mathsf{B} := \mathbf{1}\{p_V > \alpha_V\}, \qquad e_V := \Pr(\mathsf{B} = 0 \mid H_0^V).$$

Thus $\mathsf{B} = 1$ is the revision verdict and $\phi_V = \Pr(\mathsf{B} = 1 \mid (H_0^V)^c)$. The attainable level β_V belongs to this routing gate, not a terminal gate.

(i) Under H_0^V, no represented candidate is task-compatible. Querying cannot change this, so $Y_{\mathrm{Q}} = 0$ on H_0^V. This does not assert a terminal error bound conditional on the initial-inadequate, query-selected subgroup.

(ii) Escaping revision requires $p_V \le \alpha_V$. Theorem 1 therefore gives $e_V \le \beta_V$.

(iii) The disagreement event is $\{H_0^V, \mathsf{B} = 0\} \cup \{(H_0^V)^c, \mathsf{B} = 1\}$. Its probability is

$$w_V e_V + w_A \phi_V \le w_V \beta_V + w_A \phi_V.$$

This is an initial-population bound. Restricting attention to incidents that reach repair would require a guarantee for that restricted population.

(iv) Couple the policies so that the same repair on the same incident has the same supplier or query outcome, downstream selection, terminal verification, and randomness. On incidents stopped by the common early-stopping rule, set all policy outcomes to that common outcome. Otherwise, gate-based routing has outcome

$$Y_{\mathrm{route}} = \mathsf{B} Y_{\mathrm{R}} + (1 - \mathsf{B}) Y_{\mathrm{Q}}.$$

Define the failure-matched outcome by

$$Y_{\mathrm{match}} = \mathbf{1}_{H_0^V} Y_{\mathrm{R}} + \mathbf{1}_{(H_0^V)^c} Y_{\mathrm{Q}}.$$

Then

$$\mathbb{E}[Y_{\mathrm{match}}] - \mathbb{E}[Y_{\mathrm{R}}] = w_A \Delta_A, \qquad \mathbb{E}[Y_{\mathrm{match}}] - \mathbb{E}[Y_{\mathrm{Q}}] = w_V \Delta_V.$$

Let

$$L_{\mathrm{route}} := \mathbb{E}\Big[(\mathbf{1}_{H_0^V} - \mathsf{B})(Y_{\mathrm{R}} - Y_{\mathrm{Q}})\Big].$$

This is the signed shortfall $\mathbb{E}[Y_{\mathrm{match}}] - \mathbb{E}[Y_{\mathrm{route}}]$. It need not be nonnegative: the failure-matched repair knows the adequacy status, not which branch succeeds in each incident. Subtraction gives the exact identity

$$\begin{aligned} \mathbb{E}[Y_{\mathrm{route}}] - \max\{\mathbb{E}[Y_{\mathrm{R}}], \mathbb{E}[Y_{\mathrm{Q}}]\} \\ = \min\{w_V \Delta_V, w_A \Delta_A\} - L_{\mathrm{route}}. \end{aligned} \tag{17}$$

The integrand defining L_{route} is zero when the verdict agrees with adequacy and is at most one otherwise. Hence

$$L_{\mathrm{route}} \le w_V e_V + w_A \phi_V \le w_V \beta_V + w_A \phi_V,$$

which proves Equation (9). No independence between the verdict and repair outcomes is needed. A positive right-hand side is sufficient for superiority over both fixed repairs; a nonpositive one does not establish that routing is worse. The conditional repair advantages are defined when their

conditioning events have positive probability; a zero-probability type contributes zero to all weighted expressions.

Policy-specific terminal verification. If the routed policy has a different terminal gate, write $Y^{\text{actual}}_{\text{route}}$ for its outcome and retain Y_{route} for the shared-branch construction above. Define

$$\eta_{\text{term}} := \mathbb{E}[Y^{\text{actual}}_{\text{route}} - Y_{\text{route}}], \qquad \kappa_{\text{term}} := \Pr(Y^{\text{actual}}_{\text{route}} \neq Y_{\text{route}}).$$

Since the outcomes are binary, $|\eta_{\text{term}}| \leq \kappa_{\text{term}}$. The exact identity becomes

$$\begin{aligned} &\mathbb{E}[Y^{\text{actual}}_{\text{route}}] - \max\{\mathbb{E}[Y_{\text{R}}], \mathbb{E}[Y_{\text{Q}}]\} \\ &\quad = \min\{w_V \Delta_V, w_A \Delta_A\} - L_{\text{route}} + \eta_{\text{term}}, \end{aligned} \tag{18}$$

and the lower bound in Equation (9) decreases by at most κ_{term}. This separates routing losses from differences in terminal acceptance. The interpretation as a terminal-gate difference requires identical selected branches and terminal model outputs; otherwise the discrepancy includes those changes too. The gold-chosen experimental reference uses realized branch outcomes and may also use policy-specific terminal gates. It is not the failure-matched comparator in this proof.

A.10 A TOTAL-VARIATION ACTION CERTIFICATE

Suppose $0 \leq \ell(u, Y) \leq B$ and let $d_h := \sup_{u \in \mathcal{U}} \text{TV}(P^u_h, P^u_\star)$. If $|\widehat{d}_h - d_h| \leq r$ simultaneously for all represented candidates with probability at least $1 - \delta$, then on this event

$$J_\star(u_h) - J^\star \leq 2Bd_h \leq 2B(\widehat{d}_h + r), \tag{19}$$

because for $u^\star \in \arg\min_u J_\star(u)$,

$$J_\star(u_h) - J_\star(u^\star) \;\leq\; \big|J_\star(u_h) - J_h(u_h)\big| + \big(J_h(u_h) - J_h(u^\star)\big) + \big|J_h(u^\star) - J_\star(u^\star)\big| \;\leq\; 2Bd_h,$$

where the middle term is nonpositive because u_h minimizes J_h, and each absolute term is at most Bd_h since the loss is bounded by B. Thus $2B(\widehat{d}_h + r) \leq \rho$ certifies regret at most ρ (any simultaneously valid bound may replace it in Equation (20)). For a shared action u with $|\widehat{J}_h(v) - J_h(v)| \leq \eta$ uniformly over feasible actions and retained candidates, estimated regret at most ρ under every retained candidate gives $J_\star(u) - J^\star \leq \rho + 2\eta + 2Bd_h$ for any retained h. An actual-regret bound therefore needs at least one retained candidate with controlled action-conditional error, and compatibility on other queries alone does not supply it.

B CAVEAT IMPLEMENTATION AND CALIBRATION DETAILS

B.1 FORMAL TRIAGE AND DOWNSTREAM ACTION

Algorithm 1: CAVEAT for one incident.
Input: O, $\mathcal{H}(O)$, $\widehat{h}$, candidate predictions, frozen gate statistics and calibration scores, feasible queries.
1. Verify: compute discrepancies, A_V, A_S, and the ranks p_V, p_S.
2. Revise: if $p_V > \alpha_V$, withhold commitment and return REJECT/REVISE with $C(O)$.
3. Resolve: otherwise form $S(O)$; if $p_S \leq \alpha_S$ and $S(O) = \{\widehat{h}\}$ return COMMIT; if $|S(O)| > 1$ and a separating query is feasible return QUERY; otherwise return UNCERTAINTY with $S(O)$.
4. Reverify: a revised space or an acquired outcome re-enters step 1 under the prespecified update policy.
Action guard: an unresolved set authorizes an action only if its estimated regret is at most ρ under every retained candidate (Equation (20), Appendix A.10).

The returned decision is written $\pi_{\text{CAV}}(O)$. When the active space has a single member, the alternative minimum in Equation (3) is taken as $+\infty$, so $A_S = -\infty$. If the selection-null calibration scores are finite, then $p_S = 1/(n_S + 1)$; the selection gate passes only when this is at most α_S.

Downstream action. If $\widehat{J}_h(u; O)$ estimates the harm of action $u \in \mathcal{U}$ under candidate h, CAVEAT authorizes u only if

$$\widehat{\mathcal{R}}_h(u; O) := \widehat{J}_h(u; O) - \min_{v \in \mathcal{U}} \widehat{J}_h(v; O), \qquad \max_{h \in S(O)} \widehat{\mathcal{R}}_h(u; O) \leq \rho, \tag{20}$$

where $\rho \geq 0$ is the maximum tolerated regret: action without unique identity is permitted only when it is near-optimal under every retained candidate. An uncertainty set is an equivalence set only when

its members predict identical laws on every declared query,

$$P_{h,q}(\cdot \mid O) = P_{h',q}(\cdot \mid O) \quad \forall\, h, h' \in S(O),\ \forall\, q \in \mathcal{Q}. \tag{21}$$

B.2 Estimating incident consequences

The empirical discrepancy of Stage I is

$$\widehat{r}_k(h; O) := d_k\Big(P_{h,q_k}(\cdot \mid O), \widehat{P}^{\star}_{q_k}(\cdot \mid O)\Big), \qquad \widehat{\Delta}_{\mathcal{Q}}(h; O) := \max_{k \le K} \widehat{r}_k(h; O). \tag{22}$$

The estimator $\widehat{P}^{\star}_{q_k}$ depends on the modality; any discrepancy with orientation and scaling fixed before test evaluation is admissible, and observational effect estimates require stated identification assumptions.

Assumption 2 (Evidence consistency)**.** *For every $q_k \in \mathcal{Q}$, $\widehat{P}^{\star}_{q_k} \xrightarrow{p} P^{\star}_{q_k}$, and d_k is continuous in its second argument.*

Then $\widehat{r}_k(h; O) \xrightarrow{p} r_k(h; O)$ for every query and, $\mathcal{Q}$ being finite, $\widehat{\Delta}_{\mathcal{Q}}(h; O) \xrightarrow{p} \Delta_{\mathcal{Q}}(h; O)$; this does not make inadequacy detectable when competing mechanisms induce the same observable consequences (Section 4).

B.3 Finite-sample calibration of the uncertainty set

Let $\{O_i^{\mathrm{set}}\}_{i=1}^{n_{\mathrm{set}}}$ be test-disjoint calibration incidents with designated task-compatible candidates $h_i^{\dagger} \in \mathcal{A}_{\epsilon}(O_i^{\mathrm{set}}, \mathcal{Q})$, and let

$$G_i := \widehat{\Delta}_{\mathcal{Q}}(h_i^{\dagger}; O_i^{\mathrm{set}}) - \min_{h \in \mathcal{H}(O_i^{\mathrm{set}})} \widehat{\Delta}_{\mathcal{Q}}(h; O_i^{\mathrm{set}}), \qquad k_{\mathrm{set}} := \lceil (n_{\mathrm{set}} + 1)(1 - \alpha_{\mathrm{set}}) \rceil, \tag{23}$$

with $G_{(1)} \le \cdots \le G_{(n_{\mathrm{set}})}$ the ordered gaps and $\tau_{\mathrm{set}} := G_{(k_{\mathrm{set}})}$ when $k_{\mathrm{set}} \le n_{\mathrm{set}}$, otherwise $+\infty$. Under exchangeability of the calibration gaps and the test gap, the split-conformal rank argument gives $\Pr(h^{\dagger}(O) \in S(O)) \ge 1 - \alpha_{\mathrm{set}}$. The guarantee is marginal over the set-calibration population and concerns inclusion of a compatible candidate: $S(O)$ is defined for every incident by Equation (5) and returned when Stage II passes, and the bound is not conditional on that outcome. It does not imply that every member of $S(O)$ is compatible or that members are equivalent (Equation (21)).

B.4 Calibration of adaptive and multistage pipelines

Proposal, retrieval, and acquisition may adapt to the observations within an incident. CAVEAT treats the complete prespecified procedure as one map

$$\Phi : O \longmapsto \Big(\mathcal{H}(O), \widehat{h}(O), \widehat{\Delta}_{\mathcal{Q}}(\cdot; O), S(O), q^{+}(O), \pi_{\mathrm{CAV}}(O)\Big), \tag{24}$$

including prompts, retrieval depth, candidate filters, discrepancy functions, query policy, stopping rule, and tie-breaking. Parameters may be developed on separate data, but Φ is frozen before final null calibration and applied identically to every calibration and test incident. Writing $\widetilde{A}_V = \Phi_V(O)$ and $\widetilde{A}_S = \Phi_S(O)$ for the scalar statistics returned by the complete procedure and substituting them in Equation (4) gives the same finite-sample guarantees under null exchangeability; this is Mondrian, label-conditional calibration with the failure type as the label. If intermediate calibrated gates determine later queries or stopping, development, policy, final-calibration, and test data must be separate ($\mathcal{D}_{\mathrm{dev}}, \mathcal{D}_{\mathrm{policy}}, \mathcal{D}_{\mathrm{cal}}, \mathcal{D}_{\mathrm{test}}$): the policy split fixes intermediate thresholds and rules, and the final split calibrates the frozen end-to-end map. Any later change to proposer, retrieval, score, query policy, or stopping rule defines a new map and requires recalibration.

B.5 Hypothesis-space revision certificate

When the space gate fails, CAVEAT returns a certificate rather than a replacement cause,

$$C(O) := \big(p_V(O),\ \{\widehat{r}_k(h; O)\}_{h,k},\ \mathcal{H}_{\mathrm{near}}(O),\ \mathcal{Q}_{\mathrm{fail}}(O)\big), \tag{25}$$

with $\mathcal{H}_{\mathrm{near}}(O)$ the closest represented candidates and $\mathcal{Q}_{\mathrm{fail}}(O)$ the queries with the largest unexplained discrepancies: why the space lacks support and which consequences a revised proposer,

retriever, specialist, or reviewer should address. A revised space changes $\mathcal{H}(O)$ and an acquired outcome augments O; either re-enters Stage I, so an unverified replacement is never treated as compatible.

B.6 When CAVEAT applies

CAVEAT needs candidates with predicted consequences for the declared queries, incident evidence to score them against, outcome-labelled historical incidents (disjoint from test) that label the two nulls, and optionally follow-up measurements. The Chamber and CRISPRi screens supply all four with interventional outcomes and CLadder with graph-implied probes; LiveClin is the boundary case, where the cross-encoder supplies a support score rather than a predicted consequence, so the space gate is a calibrated test of that score and Appendix D.6 documents the resulting limits on exact commitment. CAVEAT does not apply when candidates make no checkable prediction, when no outcome-labelled history exists, or after any pipeline change without recalibration. The verifier costs $|\mathcal{H}(O)| \cdot |\mathcal{Q}|$ discrepancy evaluations and two rank lookups per incident, plus one utility evaluation per feasible query; rank-max adds augmented marginal and joint ranks.

Reference package and agent integration. The reference package exposes CAVEAT as a verification and repair interface: an application supplies candidates, predicted consequences, evidence, and a frozen workflow; the package returns structured decisions and diagnostic reports. `01_reference_package/README.md` specifies the integration contract, calibration requirements, and validation commands. `CAVEAT_Agent_Integration.ipynb` provides an executable bounded-workflow example and a separate replay of frozen LLM-study records, without new model calls. These illustrate integration and check implementation behavior; they do not establish deployment performance.

C Additional experimental details

C.1 Evidence ladder and independent units

Table 5 records the role and independent unit of each experiment. Controlled omissions are created after the generating mechanism or the intervention label is fixed. We remove the operative family from the active space while maintaining its observed consequences.

Compatibility in each experiment. We define $(P_{h,q}, d_q, \text{compatibility})$ for each task as follows. For SCM, we use the interventional response, squared error, and ϵ-agreement. For Chamber, we use the sensor profile, standardized distance, and generating family. For CRISPRi we use the 1,000-gene response, multiscale discrepancy, and generating family. For LiveClin we use the diagnostic profile, negative cross-encoder support, and recorded ICD-10 diagnosis. For CLadder, we use the graph effect, squared residual, and matching task signature. For external tasks, we use held-out labels instead of fitting ϵ. Compatibility labels follow each benchmark's fixed criterion: response-error agreement in the simulations, worst-query effect agreement in the matched-routing trials, and generating-family or recorded-diagnosis identity in the external tasks. The calibration guarantee concerns these specified labels; identity labels do not themselves certify agreement on unmeasured causal effects.

Controlled SCMs. The adaptive calibration study compares fixed, search-adaptive, and mismatched pipelines over 150 independent histories. The dual-null study uses 20 libraries with closed and open spaces and identifiable and confusing evidence. Every regime consists of 400 test incidents. In V-banks there are 200 detectable omissions and 199 indistinguishable omissions. In S banks there are 399 incompatible selections from the confusing regime. This is a stratified stress panel rather than an IID deployment mixture. Both gates decide upon commitment. If the non-commitment has minimum discrepancy greater than $\max\{2.5t_V,\ t_V + .08\}$, then noncommitment requests revision. All other noncommitments request a query. t_V is the archived cutoff. The second factorial experiment is conducted for 20 worlds and varies the sample size, space, discrepancy, candidate separation, action geometry, and shift. The harm tolerance is $\rho = .10$. The audit allows 722 actions among 1,440 favorable incidents. Out of those, 18 incidents have two candidates. The audit allows 1,332 actions among 86,400 factorial incidents. Out of those, 610 incidents have two

candidates. No incident exceeds the harm tolerance. Forced action is higher than the tolerance in 10,780 incidents or 12.5 percent. Consequences are the intervention response vector. The metric d_q is squared response error. Compatibility is full query mean squared error not greater than ϵ^2.

Table 5: Original frozen experimental settings; the 240-world matched-repair trial is described in the text of this section. "Controlled masking" retains the physical outcome while hiding the generating intervention family from the initial active hypothesis space.

Setting	Unit	Scale	Omission	Role
Controlled SCM	simulated incident	20 seeds; 399/null	constructed	calibration, dual nulls
Factorized SCM	simulated incident	20 worlds	constructed	cause/action regret
Causal Chamber	intervention family	29 families	controlled masking	set and query power
Jurkat CRISPRi	perturbation family	180 families	constructed	sealed biological transfer
HepG2 CRISPRi	perturbation family	180 families	constructed	sealed biological transfer
LiveClin	clinical case	167 cal.; 171 test	natural/active-space	clinical triage
CLadder + Qwen	causal world	120 test worlds	constructed	feedback and reverification

Fresh score-design comparison. We draw an independent library of eight candidate responses for each incident. We use four equally likely regimes that cross represented or omitted mechanisms with identifiable or confusing evidence. We define compatibility as full-vector RMS error at most .40. We give the scorer only the lowest observed mean squared error and the gap to the second-best candidate. We log transform, then standardize these quadratic features, and also fit class-balanced logistic models. We use separate random streams for 12,000 fitting incidents, 499 development-calibration incidents for each null, and 4,000 validation incidents.

During development, we compare the original score with learned space scores, learned selection scores, and their combination. We choose the policy with the highest retention subject to validation false acceptance of at most .10. We then freeze the selected policy before independent final calibration. We freeze the calibration thresholds before generating test incidents on a separate random stream. Each null bank uses the same regime mixture conditional on that null.

The notebook defines logits so that larger values indicate more support. We negate them to match the lower-tail $\widetilde{A}_j$ convention. Both arms commit only when both gates pass and the observed best candidate is unique. We do not use queries, revisions, or a separate calibrated set. We calibrate at the mixture level. Under the learned policy 8 of 13 incompatible selections in the identifiable subgroup are accepted. The notebook provides the data generator, fixed optimizer, subgroup results, and the paired bootstrap with 4,000 replicates.

Fresh joint-score check. A *rank-max* variant calibrates the larger of the two ranks as one joint score, either within each null or on their union. Table 6 uses a separate fresh run of the same generator. We use 20 repetitions. Each repetition includes 1,999 IID failures sampled conditional on incompatibility and 600 test incidents. We use a shared bank, which allow the V and S counts to vary naturally, and keep the learned-space parameters fixed from development. For rank-max, we add the test incident to the bank. We compute both weak marginal ranks for every point, take their maximum, and then rank the test maximum. Separate rank-max does this within each null and requires both ranks at most .10, with the own-null rank bound following from permutation equivariance and no independence needed between gates. All policies commit only when the observed best candidate is unique. We do not use revision or query. The prespecified contrast, separate rank-max minus learned space, is -16.0 retention points $[-17.3, -14.8]$ with fewer wrong commitments. We also obtain intervals resample the 20 repetitions jointly across policies (5,000 replicates). The policies use the same nominal levels, but their observed error rates can differ.

Table 6: **Fresh synthetic comparison.** Retention; V/S/mixture acceptance (%); wrong counts/12,000. Cases: 4,593 compatible, 5,988 V, 1,419 S. All 20 repetitions use 1,999 shared calibration failures at .10; all but pooled calibrate per null.

Rule	Retention	Wrong	V	S	Mixture
Original separate	15.39	17	0.23	0.21	0.23
Learned space	68.10	455	5.93	7.05	6.14
Pooled rank-max	67.28	757	10.42	9.37	10.22
Separate rank-max	52.10	156	2.25	1.48	2.11

Randomized physical interventions. The light tunnel provides randomized interventions, nine sensors, and an empirical DAG (Gamella et al., 2025). We rotate 29 target families through known, calibration-unknown, and test-unknown roles. After rejection we search a 14-family complement

Table 7: **Fresh verification-only study at levels .10/.10.** Counts (%) [Wilson 95% intervals], conditional on compatible/incompatible selection. Both arms share selections and evidence; intervals condition on frozen fitted and calibrated rules.

Verifier	Retention	False acceptance
Original scores	642/4,659 (13.78) [12.82, 14.80]	16/7,341 (0.22) [0.13, 0.35]
Learned space score	3,132/4,659 (67.22) [65.86, 68.56]	432/7,341 (5.88) [5.37, 6.45]

using descendant fingerprints. This setup tests structural-complement retrieval over a finite atlas rather than ontology-free discovery. We preregister a three-step sensor ladder with three initial sensors, one selected sensor, and all nine sensors. We define consequences as memory-split sensor means. We measure discrepancy as mean squared sensor error standardized by memory variance. The retriever compares each observed anomaly with the descendant fingerprint of each target. The wind tunnel (`wt_validate_v1`; 21 families, 11 sensors, roles 11/5/5 per rotation, nominal $\alpha = 0.20$, attainable $1/6$) was run once, sealed, under the frozen light-tunnel design.

Biological perturbations. The source study reports 262,956 Jurkat and 145,473 HepG2 cells (Nadig et al., 2025). We use 180 eligible intervention families and 1,000 response genes, with family-level splits so that no family crosses calibration and test. Dense and isolated omissions vary null composition. Protocols, budgets, thresholds, and checksums were frozen before each sealed evaluation.

Each consequence is a memory-split pseudobulk profile. We compute discrepancy across 200- and 1,000-gene views, combining magnitude error, cosine alignment, sign agreement, and strong sign contradiction with fixed coefficients $1, .25, .15$, and $.20$. Table 8 reports retention and wrong commitment with both gates. HepG2 retains 27 of 89 compatible selections, with 6 wrong commitments over 360 test units. Jurkat retains 22 of 116, with 8 wrong commitments over 720 units.

Table 8: **Biological retention with both gates.** Dense-neighbour omission, matched calibration, and 1,000-gene profiles. Retention is conditional on an initially compatible selection; wrong commitment uses all test units.

Setting	Test units	Initially compatible	Retained	Retention (%)	Wrong commitments
HepG2	360	89	27/89	30.3	6/360
Jurkat	720	116	22/116	19.0	8/720

LiveClin. Of 1,407 cases, 727 form the development atlas, 342 are used during protocol development, and 338 untouched cases split into 167 calibration and 171 test. The pool holds 445 exact diagnoses, 72 level-2 categories, and 16 chapters. The frozen cross-encoder observes the initial case and may request up to three further recorded evidence items; $P_{h,q}$ is a diagnosis description with a development-case prototype and d_q the negative support after evidence q is revealed (Wang et al., 2026). At stage t the active space is the top 10 of the 445 codes; with recorded diagnosis y_i as delayed reference, never given to retrieval, scoring, querying, or the decision rule, the stage is V-null when y_i is absent, selection-null when present but not selected, and compatible when selected. Sequential calibration uses one trajectory-level score per case and null (the least favorable V-null score across stages and, separately, the least favorable selection-null score), so a case can enter both banks at different stages while the nulls stay disjoint at any one stage: 139 cases supply a space-null score and 42 a selection-null score, 19 both. With $\alpha_V = \alpha_S = .05$, attainable levels are $\beta_V = .050$ and $\beta_S = .047$; the same thresholds govern every stage with ties deferred, Theorem 1 bounds each null's wrong commitment at any stage, and the union bound gives a per-case bound of $\alpha_V + \alpha_S = .10$. Two audited artifacts are reported rather than corrected: exam-style stems revealed as evidence name the recorded diagnosis in 1.9% of development cases (0.6% of stems), and 512-token truncation removed at least one revealed item in 10.6% of cases, on the same pipeline used for calibration.

CLadder feedback. We evaluate a controlled extension of CLadder in which an LLM receives a candidate set of causal graphs and selects one (Jin et al., 2023; Yang et al., 2024). For a probe $q = (A, B)$, $P_{h,q}$ is the total effect predicted by graph h, and d_q is its squared residual from the

Table 9: **Feedback policies on the 120-world CLadder extension.** Qwen2.5-7B-Instruct; entries are percentages of all 120 test worlds, with exact 95% binomial intervals below. Task-correct counts compatible selections whether or not they receive commitment; recovered counts correct commitments from initially incompatible selections. Expansion and query policies reverify the final selection. The report arm additionally exposes the current-state residual report. The oracle uses gold information to choose a probe, not the final answer.

Policy	Wrong commitment	Correct commitment	Task-correct	Recovered
Raw Qwen	75.8 [67.2, 83.2]	24.2 [16.8, 32.8]	24.2 [16.8, 32.8]	0.0 [0.0, 3.0]
Self-reflection	75.8 [67.2, 83.2]	24.2 [16.8, 32.8]	24.2 [16.8, 32.8]	0.0 [0.0, 3.0]
Expansion only	0.0 [0.0, 3.0]	32.5 [24.2, 41.7]	38.3 [29.6, 47.6]	13.3 [7.8, 20.7]
Random query	0.8 [0.02, 4.6]	33.3 [25.0, 42.5]	40.0 [31.2, 49.3]	14.2 [8.5, 21.7]
CAVEAT query	0.8 [0.02, 4.6]	35.0 [26.5, 44.2]	40.0 [31.2, 49.3]	14.2 [8.5, 21.7]
Oracle query	0.8 [0.02, 4.6]	33.3 [25.0, 42.5]	37.5 [28.8, 46.8]	13.3 [7.8, 20.7]
CAVEAT query + report	6.7 [2.9, 12.7]	52.5 [43.2, 61.7]	59.2 [49.8, 68.0]	31.7 [23.5, 40.8]

observed outcome. Task compatibility requires matching the ancestry, direct-edge, and observed-common-parent signature of the true graph.

The feedback policies allow one repair round and at most one additional model call. To fit this budget, CAVEAT uses its typed certificate to plan the candidate update and query together, without re-testing the updated space before querying. The model then makes one new selection, which is reverified at both gates. We compare raw selection, self-reflection, expansion, random-query, CAVEAT-query, oracle-query, and residual-report policies. Each is a complete observation-to-action map, with its own space, selection, and set thresholds from the same disjoint 120-world terminal-calibration split; an additional 80-world split calibrates the initial checks. Raw selection and self-reflection are forced-choice controls. Prompts, candidate identifiers, parsing rules, and thresholds are frozen before evaluation on the 120 test worlds. Outputs without a valid candidate identifier never commit, are excluded from the selection-null bank, and are counted in the archive.

The primary reasoner is Qwen2.5-7B-Instruct, using 4-bit NF4 quantization and FP16 computation on one T4 in Phase 11C or one A100 in Phase 14A. Decoding is greedy, with batch size 4, a 3,000-token input cap, at most 112 new tokens, and seed 20270830.

Table 9 compares the feedback policies. The initial candidate set is adequate in 60 of 120 worlds; expansion raises this to 89. Self-reflection changes one selected identifier but no task-level outcome. With the CAVEAT query and reverification, 42 answers receive correct commitment, including 17 recovered from initially incompatible selections, while one commitment is wrong. The residual report increases correct commitments to 63 but also raises wrong commitments to eight.

Table 10 repeats the comparison across six reasoners, each calibrated as its own frozen pipeline. Raw wrong commitments range from 87 to 98 of 120 worlds. The CAVEAT query policy reduces these to one to seven, with 22–57 correct commitments. The table also reports wrong commitments within the 31 inadequate final spaces and after adding the residual report.

Matched one-step repair: design-stage trial. Version 1 (640 earlier model IDs excluded; 80 routing-calibration, 320 final-calibration, 240 test, 20 smoke-test worlds; Qwen2.5-7B) gives routing 51/1 correct/wrong commitments, versus 50/1 for always-revise, 30/4 for always-query, and 42/1 for random repair. The primary random contrast is 3.75 points [0.00,7.50] (95% paired bootstrap; exact paired $p = .078$) and the always-revise contrast 0.42 [-2.50,3.33]: neither advantage is established. Routing uses 203 revisions and 22 queries, retains 24/52 initially compatible answers, and like always-revise recovers 21/53 repairable omissions (query-only none); all policies use 1.9375 calls per world. Five informative base probes left little ambiguity for a query, and qualitative compatibility differed from quantitative fit.

Table 10: **CLadder feedback across six reasoners.** Each reasoner is calibrated as its own frozen pipeline on the same splits. Correct/wrong entries are counts out of 120 test worlds. The query policy's wrong-commitment percentage is shown with its exact 95% interval. The V-null column counts wrong commitments among the 31 worlds whose final candidate sets contain no task-compatible explanation; these are included in, not added to, the total wrong counts. The final column gives wrong commitments out of 120 when the residual report is added. The attainable β_V is 9.4% for all six reasoners. Qwen models are Qwen2.5-Instruct; Phi-3.5-mini is the instruct model.

Reasoner	Raw correct/wrong	CAVEAT query correct/wrong	Query wrong (%)	V-null wrong	+ report wrong
Qwen2.5-3B 3B	25/95	22/2	1.7 [0.2, 5.9]	2/31	5
Phi-3.5-mini 3.8B	22/98	29/2	1.7 [0.2, 5.9]	0/31	2
Qwen2.5-7B 7B	29/91	42/1	0.8 [0.02, 4.6]	1/31	8
Qwen2.5-14B 14B	32/88	47/7	5.8 [2.4, 11.6]	3/31	8
phi-4 14B	33/87	45/2	1.7 [0.2, 5.9]	1/31	8
Qwen2.5-32B 32B	33/87	57/3	2.5 [0.5, 7.1]	2/31	7

Version 2: complementary repair opportunities. The new protocol excludes those 640 IDs and all 660 version-1 worlds; Qwen2.5-7B and 14B share 100 routing-calibration, 320 final-calibration, 240 test, and 20 smoke-test worlds (NF4 with BF16 computation on an A100, greedy batch-one decoding, at most 32 new tokens). Task queries have outcome Y or intervened variable X; compatibility requires worst task-effect gap at most .02; scoring uses mean squared residual on observed probes. With three base probes and noise s.d. .03, exact-present and omission regimes use candidates separated on base evidence; ambiguous-present worlds hold a compatible graph and an incompatible neighbor that agree on the base probes but differ by at least .15 on an available probe. No task-compatible neighbor exists in this construction, so ambiguous-present replaces version 1's qualitative equivalent-present regime. The router revises when its development-calibrated space gate fails and otherwise queries, with a shared early stop for a supported unique selection; revision supplies four candidates (the truth, under repairable omission) and no measurement, a query keeps the candidates and buys the probe with the largest predicted disagreement weighted by fit to existing evidence. The four policies share initial selections, eligibility, and at most one further model call, and both repairs are recorded for every eligible world. Each complete policy receives separate terminal V/S banks at .10/.10 without filtering S by the V gate and requires a uniquely best verified selection; all banks exceed minimum support; token, measurement, and candidate costs are reported separately.

Table 11: **The failure determines which repair helps.** Correct commitments per regime and over all 240 worlds. Cause present: the true cause is in the space, with clear (61 worlds) or ambiguous (57) evidence. Cause missing: it is absent, and revision can (repairable, 64) or cannot (unrepairable, 58) supply it. Gain: CAVEAT minus the policy, in % of the 240 worlds [95% paired world-bootstrap interval]. Every arm makes 3–5 wrong commitments. Gold-chosen: the better repair outcome in each world, using the true answer.

	Qwen2.5-7B						Qwen2.5-14B					
	Cause present		Cause missing				Cause present		Cause missing			
Policy	clear (61)	ambig. (57)	repairable (64)	unrepairable (58)	all 240	Gain [95%]	clear (61)	ambig. (57)	repairable (64)	unrepairable (58)	all 240	Gain [95%]
CAVEAT routing	35	31	20	0	86	—	48	39	30	0	117	—
Always revise	36	0	20	0	56	12.5 [7.5, 17.5]	49	0	32	0	81	15.0 [9.6, 20.4]
Always query	35	31	0	0	66	8.3 [5.0, 12.1]	48	39	0	0	87	12.5 [8.3, 16.7]
Random repair	36	20	9	0	65	8.8 [5.0, 12.9]	49	23	14	0	86	12.9 [8.3, 17.5]
Gold-chosen repair	40	31	20	0	91	—	53	39	32	0	124	—

Routing exceeds each comparator at both model sizes (all Holm-adjusted paired $p < 10^{-4}$); the random-policy gain is 8.8 [5.0,12.9] points at 7B and 12.9 [8.3,17.5] at 14B. World-bootstrap intervals condition on the fitted calibration; intervals resampling the 26 story clusters are also positive; the shared worlds across models are not independent replications. Routing queries all 57 ambiguous worlds and revises 59 of 64 repairable omissions in both runs (the rest are early stops and queries). The gold-chosen repair reaches 91 and 124 against routing's 86 and 117; it uses gold outcomes and the fixed policies' terminal gates, whereas routing has its own calibration. For Proposition 2, the

Table 12: **Matched repair, version 2.** Correct and wrong commitments out of 240; retained answers over initially compatible selections. Gain is CAVEAT minus the row's policy in percentage points [95% paired world-bootstrap interval]. The gold-chosen repair takes the better realized outcome of the two fixed repairs in each world, using gold labels and their own terminal gates.

Policy	Correct	Wrong	Retained	CAVEAT gain [95% interval]
Qwen2.5-14B				
CAVEAT routing	117	5	75/77	—
Always revise	81	5	44/77	15.0 [9.6, 20.4]
Always query	87	5	75/77	12.5 [8.3, 16.7]
Random repair	86	5	61/77	12.9 [8.3, 17.5]
Gold-chosen repair	124	5	75/77	—
Qwen2.5-7B				
CAVEAT routing	86	4	56/57	—
Always revise	56	4	30/57	12.5 [7.5, 17.5]
Always query	66	4	56/57	8.3 [5.0, 12.1]
Random repair	65	3	47/57	8.8 [5.0, 12.9]
Gold-chosen repair	91	4	56/57	—

space verdict disagrees with the failure type in 9/122 inadequate worlds (queries and early stops) and 0/118 adequate worlds for both models; 9/122 is 7.4%, compared with the marginal attainable level 0.093 of the routing bank. The present-truth regimes contribute 5 and 5 of the shortfall against the gold-chosen repair, where re-selection over the expanded space succeeded and the probe did not: the reasoner, not the verdict. For Proposition 2(iv), the empirical estimates of $w_V\Delta_V$ and $w_A\Delta_A$ are .083 and .125 at 7B, and .133 and .158 at 14B. Using the routing bank's attainable level $\beta_V \approx .093$ and the observed false-rejection rate of zero, the plug-in right-hand side of Equation (9) is 3.6 and 8.6 percentage points, compared with observed gains of 8.3 and 12.5 points over the stronger fixed repair. The shared-branch reconstruction agrees with routing's correct-commitment outcome on every test world, giving 86 and 117 correct commitments and no observed terminal-outcome mismatches (`reproduction/routing_value_identity.py` in the artifact). These plug-in evaluations are not confidence bounds. Version 2 changes population, queries, compatibility criterion, and query rule, so its retention change from version 1 does not isolate a scoring effect.

C.2 Imperfect candidate supply across reasoners

We extend the matched-repair protocol to Qwen2.5-3B, 7B, 14B, and 32B, Phi-3.5-mini, and OLMo-2-1124-13B. All six share a new panel of 100 routing-calibration, 320 final-calibration, 240 test, and 20 smoke-test worlds, disjoint from the version-2 panel. The test split contains 73 exact-present, 61 ambiguous-present, 50 repairable-omission, and 56 unrepairable-omission worlds. Revision supplies a compatible candidate in originally repairable omissions with probability $r \in \{0, .25, .50, .75, 1\}$. Failed supply replaces that candidate with an incompatible one, leaving the initial candidates and evidence unchanged. The router observes neither the regime nor the supply coin, and failed repairs remain in the denominator.

The primary contrast is routing minus random repair at $r = .50$; the additional models extend the original 7B plan. Each policy and supply condition receives matched terminal V/S calibration at .10/.10. The comparison therefore tests imperfect supply with recalibration, rather than transfer of frozen thresholds to an unknown supply rate. The 240 test worlds and 26 stories are shared, not independent replications across models. Intervals resample paired story clusters and condition on fitted calibration. Query quality remains as in the controlled protocol. Qwen uses strict parsing and at most 32 new tokens; Phi and OLMo use a lenient parser and at most 256. Token and measurement costs differ across repair types and are reported separately. The reproduction notebook regenerates counts, comparisons, and supply curves from the six frozen archives.

Figure 3 shows the full supply sweep.

Post-hoc selection audit. At full supply, we recompute each candidate's discrepancy from the archived graphs and the terminal evidence supplied to each reasoner in the routed arm, without new model calls. We compare the unique discrepancy minimizer with the reasoner's selection. The minimizer is task-compatible in 178 of 240 worlds for each reasoner, whereas the reasoner's selection

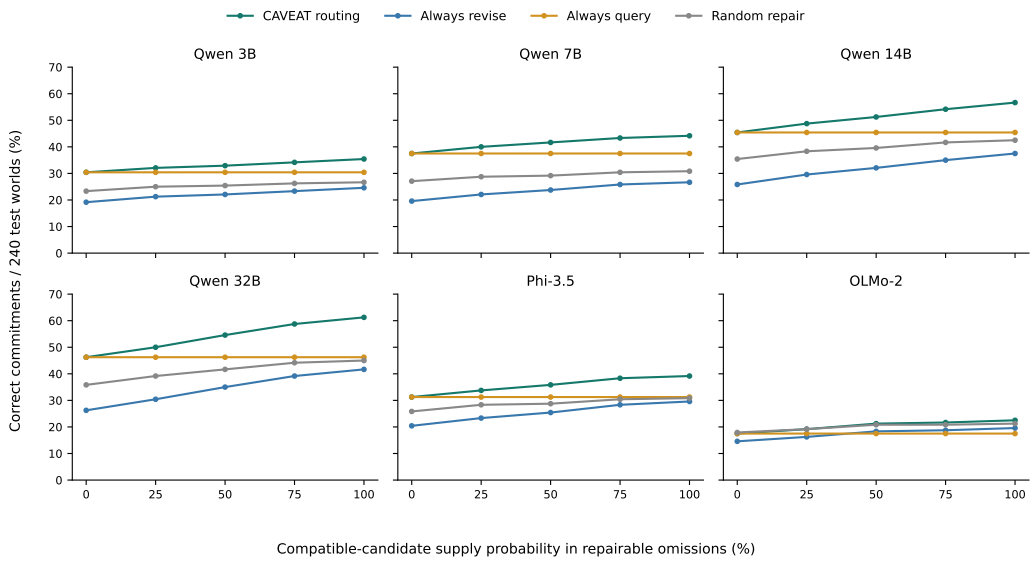

Figure 3: **Routing under imperfect candidate supply.** Correct commitments as a percentage of all 240 test worlds, including noncommitments, as the probability that revision supplies a compatible candidate in repairable omissions increases. Each policy and supply condition receives matched terminal calibration at $\alpha_V = \alpha_S = .10$. All six reasoners share the test worlds. Curves show observed rates without uncertainty intervals.

is compatible in 54–147 worlds. The 178 comprise 73 exact-present, 61 ambiguous-present, and 44 repairable omission worlds. The ambiguous pairs were constructed to be separated by the available probe. These are candidate selection counts, not calibrated commitment counts; the audit does not evaluate a newly calibrated reselection policy.

C.3 Metrics and baselines

Let N contain test incidents, $\mathcal{A}_i := \mathcal{A}_\epsilon(O_i, \mathcal{Q})$ their task-compatible candidates, and $\mathsf{C}_i \in \{0,1\}$ indicate commitment to $\widehat{h}_i$. Safety and retained usefulness are measured by $\text{Wrong} = |N|^{-1} \sum_{i \in N} \mathsf{C}_i \mathbf{1}\{\widehat{h}_i \notin \mathcal{A}_i\}$ and $\text{Correct} = |N|^{-1} \sum_{i \in N} \mathsf{C}_i \mathbf{1}\{\widehat{h}_i \in \mathcal{A}_i\}$; both denominators include noncommitments, and "under omission" restricts N to inadequate-space incidents. For fixed selections, *retention* and *false acceptance* are commitment rates conditional on compatible and incompatible selection, respectively; for LLM repair workflows, retention is the fraction of initially compatible selections that end in a correct commitment, and recovery the fraction of initially incompatible ones that do. Chamber recovery instead measures correct retrieval among omitted roles, before commitment. Inadequacy detection on V-null incidents is $\mathbf{1}\{\mathsf{Reject}/\mathsf{Revise}_i\}$, with query and uncertainty reported separately since they prevent commitment without claiming localization; resolution metrics are truth-in-set coverage, set cardinality, exact localization, query rate, action coverage, and actions whose loss exceeds ρ. Baselines share the frozen scorer and change only the decision interface. Forced choice emits the top candidate, V-only omits the selection gate, margin and entropy cutoffs use relative-score heuristics, one-class rejection uses one absolute inadequacy score, and split conformal commits only for a singleton prediction set. V-only isolates the contribution of selection verification within our framework; it is not an implementation of Li et al. (2026). Unless otherwise specified, intervals are exact binomial for patient, world, and simulated-incident rates; history-level Student-t, Bonferroni-adjusted over six pipeline regimes, for adaptive SCMs; family bootstraps for Chamber and CRISPRi; exact McNemar for paired LiveClin and CLadder comparisons. Cells, repeated measurements, and prompt variants are never independent replicates.

Table 13: **Calibration levels, guarantees, and observed rates.** Bounds apply under the stated assumptions; nominal levels and descriptive comparisons are identified separately. Observed rates are finite-sample estimates, not deterministic checks of the bounds. LiveClin uses $\alpha_V = \alpha_S = .05$ and a 10% per-case union bound. The operating-region ceiling $\min\{1, \beta + \mathrm{TV}(P_1^F, P_0^F)\}$ is marginal, not a bound for each displayed regime. Sources: Tables 4, 17, 6, 7, 16; Figure 4; Appendix C.1.

Result	Setting	What is counted	Bound or reference level	Observed
Theorem 1 + union bound	LiveClin	Wrong exact commitment per case	10% per case	0.0%
Theorem 1	K562, three levels	Wrong commitment under omission	Nominal 5% / 10% / 20%	2.6% / 8.0% / 17.2%
Prop. 2(ii)	Matched repair, 122 inadequate worlds	No revision given H_0^V	9.3%	7.4%
Lemma 1	Adaptive SCM, old / new cutoff	Wrong commitment	$\min\{1, \beta_n + d_+\}$ / nominal 10%	43.7% / 10.4%
Theorem 1	SCM, pooled / separate	V-null acceptance	No 10% V-null bound / 10%	10.4% / 2.3%
Theorem 1	Light tunnel, $n_S \leq 7$, $\alpha_S = .10$	Any commitment	0.0%	0.0%
Prop. 1	SCM, identifiable / confusing	Retention by regime	No subgroup bound	99.4% / 5.2%
Evidence acquisition	Light tunnel, 3 / 9 sensors	Recovery	Descriptive comparison	17.7% / 48.3%

D ADDITIONAL RESULTS

D.1 THE LEVEL AS A TRADE-OFF

On exposed K562 development data with matched scores, candidate pools, and calibration incidents, CAVEAT's wrong commitment under dense omission is 2.6, 8.0, and 17.2% at $\alpha = 0.05, 0.10, 0.20$, with joint exact identity 11.6, 20.2, and 30.8% (Figure 4); among calibrated open-set baselines at the same levels, Mahalanobis reaches 4.7, 14.0, and 22.9% identity, and energy and kNN detectors stay below 9%. Under isolated-decoy omission those detectors identify 55–59%, because isolated decoys are far from every family, which CAVEAT's consequence-based statistic does not exploit.

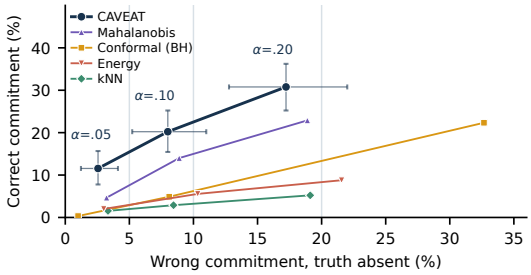

Figure 4: **CAVEAT keeps more correct commitments at every level.** K562 screens, dense-neighbour omission, 180 families; every rule shares the scorer, candidate pools, and calibration; baselines are the Mahalanobis (Lee et al., 2018), energy (Liu et al., 2020), and kNN (Sun et al., 2022) detectors and candidate-level conformal BH (Bates et al., 2023). Whiskers: 95% family-bootstrap intervals. CAVEAT exceeds Mahalanobis, the best baseline, by 6.2 percentage points [2.6, 10.2] at $\alpha = .10$.

Level sweep on the sealed LiveClin unit (post-hoc). Phase 14B re-scores the frozen LiveClin pipeline at total $\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.30\}$, with $\alpha_V = \alpha_S = \alpha/2$ and $\alpha_{\mathrm{set}} = \alpha$. Every threshold is recomputed from the calibration split alone and applied once to the 171 test cases. Because the outcomes were unsealed in Phase 12C, this is a post-hoc sensitivity analysis of the frozen operating point. Re-scoring reproduces every sealed action and prediction; one returned set differs by one member at the set threshold.

Table 14 shows that changing the level mainly changes the returned directions. Wrong exact commitment stays at 0/171, and correct exact commitment reaches only 1/171 from $\alpha = .10$. At $\alpha = .30$, split conformal on the same scores and candidates makes eight wrong and three correct exact com-

mitments. Its sets can shrink to singletons from lists that omit the diagnosis; the space gate refuses these commitments.

Table 14: **LiveClin level sweep (post hoc).** Correct/wrong exact commitments are counts out of 171 test cases. Total level α sets $\alpha_V = \alpha_S = \alpha/2$ and $\alpha_{\text{set}} = \alpha$. Direction coverage and mean size are conditional on returning a level-2 direction, not percentages or sizes over all 171 cases. Mean sizes are rounded to whole categories.

α	CAVEAT exact correct/wrong	Direction coverage (%)	Mean direction size	Split conformal exact correct/wrong
0.05	0/0	98.8	52	0/0
0.10	1/0	91.8	33	2/0
0.15	1/0	85.7	22	2/0
0.20	1/0	82.7	17	2/0
0.30	1/0	66.4	10	3/8

D.2 Constructed versus natural hypothesis-space nulls

Controlled studies calibrate on constructed omissions and LiveClin on natural ones. Where both are available, we can compare their score populations. Among the 342 LiveClin development cases scored by the frozen cross-encoder, 147 naturally omit the recorded diagnosis from the top-10 candidates. For each of the other 195 cases, we construct a space null by withholding the recorded diagnosis. The best represented score is unchanged unless the diagnosis ranked first, in which case we use the runner-up.

Table 15 compares A_V at each evidence stage and at the trajectory maximum. Mean scores range from 7.9–10.4 for natural nulls and 8.3–11.0 for constructed nulls. A one-dimensional domain classifier has AUC 0.51–0.55, with every interval covering .5; Kolmogorov–Smirnov tests do not reject equality. A cutoff calibrated on one population gives 2.1–6.8% exceedance on the other at nominal 5%. These checks describe score similarity for this scorer and benchmark. They neither establish exchangeability nor identify a direction of bias.

Table 15: **Natural and constructed LiveClin space nulls.** All rows use 147 natural and 195 constructed cases. Means are natural/constructed; AUC intervals are the archived 95% intervals. The last two columns give exceedance rates on one population using the cutoff calibrated on the other, at nominal 5%. Stage rows and the trajectory maximum reuse the same cases.

Evidence	Mean A_V natural/constr.	AUC [95% interval]	KS p	Natural at constr. cutoff (%)	Constr. at natural cutoff (%)
Stage 0	9.47/10.14	0.549 [0.484, 0.611]	0.12	4.1	4.6
Stage 1	7.87/8.38	0.533 [0.470, 0.597]	0.37	6.8	2.1
Stage 2	8.78/8.98	0.519 [0.458, 0.580]	0.55	6.1	4.1
Stage 3	8.18/8.30	0.510 [0.447, 0.574]	0.92	3.4	4.6
Trajectory max	10.42/10.95	0.538 [0.476, 0.598]	0.16	4.8	2.6

D.3 Controlled SCM panels

Table 16 reports the SCM panels. The adaptive rows compare a threshold calibrated before search with one calibrated on the adaptive pipeline over 150 independent runs. In the dual-null panels the space gate passes 1,617/8,000 indistinguishable omissions, and 6,202 further incidents are queried despite failing it under the archived noncommitment router. The Chamber rows preserve the original Phase 6D-v2 run (28.1% selected-query recovery); Table 17 uses the matched Phase 17A run (27.6%) for its random-query comparison.

Table 16: **Safety and supported resolution across controlled and sealed units.** Entries are percentages; brackets give 95% intervals over the stated units. The screen rows report the archived hypothesis-space-gate wrong-commitment rate among eligible families in the expansion regime (metric `wrong_expand`), not the all-incident dual-gate counts of Appendix C (6/360 and 8/720).

Safety			Detection, retention, and resolution		
Setting	Evaluation population	Error rate ↓	Setting	Evaluation population	Supported outcome ↑
Adaptive search: recalibrated	all; 150 histories	wrong 10.4 [9.7, 11.2]	Adaptive search: recalibrated	all; 150 histories	exact 22.9 [22.0, 23.8]
Adaptive search: threshold reused	all; 150 histories	wrong 43.7 [41.0, 46.5]	Adaptive search: threshold reused	all; 150 histories	exact 41.4 [40.3, 42.4]
Dual-null: detectable	H_0^V; 8,000 incidents	wrong 0.0 [0.0, 0.05]	Dual-null: detectable	H_0^V; 8,000 incidents	reject 99.9 [99.9, 100.0]
Dual-null: indist.	H_0^V; 8,000 incidents	wrong 0.48 [0.34, 0.65]	Dual-null: indist.	H_0^V; 8,000 incidents	query 97.3 [96.9, 97.6]
Action SCM	authorized; 20 worlds	harmful 0.0 (all worlds)	Action SCM	favorable; 20 worlds	cause/action 67.8/50.1
Chamber: 3 sensors	omitted; 29 families	wrong 8.4 [2.0, 16.3]	Chamber: 3 sensors	omitted; 29 families	set/exact 86.2/17.7
Chamber: +1 sensor	omitted; 29 families	wrong 8.4 [2.0, 16.3]	Chamber: +1 sensor	omitted; 29 families	set/exact 74.5/28.1
Chamber: all 9	omitted; 29 families	wrong 10.3 [2.0, 21.2]	Chamber: all 9	omitted; 29 families	set/exact 81.5/48.3
Jurkat CRISPRi	eligible; 180 families	wrong 10.0 [6.1, 14.2]	Jurkat CRISPRi	eligible; 180 families	proposal/exact 83.3/11.4
HepG2 CRISPRi	eligible; 180 families	wrong 7.8 [3.9, 11.7]	HepG2 CRISPRi	eligible; 180 families	proposal/exact 94.4/17.8

D.4 Exact random-query comparison on the frozen light-tunnel panel

Table 17: **Causal Chamber.** A: recovery (%) over 203 omitted roles, 29 families; random averages six sensors. B: matched-family audit at .20 (space) and .10 (selection); accuracy (%) over 435 known roles, commitments over all 638.

A. Retrieval after rejection

Evidence	Sensors	Recovery	95% interval
Initial	3	17.7	[5.9, 31.5]
Random mean	4	22.2	[10.7, 35.6]
Selected	4	27.6	[13.3, 42.9]
All sensors	9	48.3	[31.5, 66.0]

B. Matched-family calibration

Evidence	Raw accuracy	S-null families	Commits correct, wrong
Three	42.3	5–12	3, 4
All nine	76.3	0–7	0, 0

The random baseline in Figure 2B and Table 17A is the exact expectation of a uniform random query on the fixed 203-role panel, averaging the six admissible unused sensors for every incident that reaches acquisition (22.25%; the archive's single random draw gave 21.18%), with every archived outcome reproduced and candidate order, tie-breaking, gates, and the entropy selector frozen. Selected recovery is 27.59%, a paired difference of 5.34 points with a descriptive family-bootstrap 95% interval of $[-1.40, 13.96]$ (seven roles averaged per family, 29 paired records, 20,000 resamples, seed 20260909; five families favor selected queries, five random, nineteen tie). Rotated libraries overlap and replicates do not refit retrieval or calibration, so enumeration removes random-draw noise only. The 120-world LLM query comparison is binomial over test worlds conditional on each frozen arm; in its saved paired outcomes two worlds favor selected acquisition and none random (exact McNemar $p = .50$).

D.5 Sealed wind tunnel: resolution of the structural fingerprint

Table 18 reports the sealed Phase 6J run under the frozen Chamber design, with family-bootstrap 95% intervals. Among omitted families, wrong commitment falls from 100% under forced choice to 13.3% with three sensors and 7.6% with all eleven (nominal .20, attainable $1/6$). For known families, identity rises from 26.4% to 58.9%, while wrong commitment falls from 17.7% to .4%. Exact recovery of an omitted family remains low, at 1.9–5.7%.

The official graph reveals a structural limit before outcomes are read. With eleven sensors, the 21 targets have eight distinct descendant fingerprints plus six fingerprint-identical classes, including `osr_2`, `pot_2`, and `v_2`. With three sensors there are five classes, one containing eleven all-zero targets. A fingerprint-only rule cannot separate members of a class, although equal reachability does not imply equal sensor distributions.

The set-valued rule retains such alternatives, but refining the set with one scalar query removes the true member in 8.0 percentage points [2.0,16.0] of cases and lowers cost-adjusted utility by .027 [.003,.046]. Structural Hit@1 is 1.9%, against a 10% random rate. Set coverage is 66.7%, versus 42.9% for a shuffled atlas at mean size 4.92. The paired difference of 23.8 points $[-14.3, 57.1]$ does not establish an atlas advantage. Thus, the prespecified Hit@1 and query-utility gates fail and are recorded as such. Development negative controls recover within-recording identities from graph-unreached sensors, so known-family gains may reflect recording background.

Table 18: **Sealed wind-tunnel sensor ladder (Phase 6J).** Percentages with archived 95% family-bootstrap intervals; 21 families. Omitted-family and known-family metrics have different denominators. These are the original space-gate/retrieval arms, distinct from the later two-gate comparisons. Nominal space level is .20; attainable level is $1/6$. A dash indicates no returned set.

Arm	Omitted wrong	Omitted recovery	Omitted set covers	Known wrong	Known identity
Forced	100.0 [100.0, 100.0]	0.0 [0.0, 0.0]	–	26.0 [9.5, 44.2]	74.0 [55.8, 90.5]
3 sensors	13.3 [1.0, 26.7]	1.9 [0.0, 5.7]	65.0 [45.0, 85.0]	17.7 [4.3, 32.9]	26.4 [11.7, 42.4]
+1 selected sensor	13.3 [1.0, 27.6]	3.8 [0.0, 8.6]	57.0 [38.0, 76.0]	17.7 [5.2, 32.5]	26.4 [11.7, 42.9]
all 11 sensors	7.6 [1.0, 18.1]	5.7 [0.0, 15.2]	51.2 [31.0, 71.5]	0.4 [0.0, 1.3]	58.9 [41.1, 76.2]

D.6 Recovery after hypothesis-space rejection (post-hoc)

We performed two post-hoc audits to examine whether an earlier rejection could later lead to a correct commitment. The first used a second retrieval step based on a historical revision corpus containing 131 ICD codes. None of the three retrieval methods (i.e., certificate-guided, naive, random-item retrieval) produced an exact commitment. None produced a wrong commitment (0/171). A major constraint was that only 12 of the 76 missing diagnoses were represented in this corpus. Even when the correct diagnosis was retrieved, the median candidate-set size remained 12. There are still too many alternatives for the verifier to isolate a single diagnosis.

The second audit removed this restriction by searching the full WHO ICD-10 index. It also allows decisions at three levels. They are exact diagnosis, level-2 category, and chapter. Each level had its own dual-null calibration. Revision was applied only to cases without an initial sealed commitment, and ten candidates were retrieved for each case. The revision query was selected using the 727-case development set with leave-one-case-out simulated omission. On this development set, the residual-conditioned query retrieved the exact diagnosis in 7 of 150 cases, compared with 3 of 150 for the whole-case query.

On the test set, no wrong commitments occurred at either the exact or level-2 level, and the single initial exact commitment was retained. The selected query added the missing diagnosis to the candidate set in 3 of 76 cases (3.9%, 0.8–11.1), compared with 2 of 76 for the whole-case query ($p = 1$). The superior performance identified during development thus did not reproduce in testing. Revision increased level-2 containment from 47/76 to 53/76 cases (69.7%, 58.1–79.8) and retrieved 41 of 94 atlas diagnoses that had been missed by the initial top-10 retrieval. No level-2 commitments were made in any arm. One arm produced three chapter-level commitments, two of which were correct, but the number of chapter-level null cases was too small to support a reliable conclusion.

We also audited the fixed verifier by adding the recorded diagnosis to every test case in which it was missing and then applying the calibrated thresholds. Even with the correct diagnosis included, the verifier produced no exact singleton among 171 cases and only four level-2 singletons (2.3%, 0.6–5.9); the gates passed in only one case. The median gap between the best and second-best candidate scores was 1.4 at the exact level and 1.9 at level 2, both well below the corresponding selection thresholds of 5.3–8.7 and 5.4–6.3. This suggests that the scorer did not separate the correct diagnosis clearly enough from nearby alternatives. It does not imply that the underlying population evidence distributions are identical; Proposition 1 concerns that population-level setting.

Across these audits, changes to candidate coverage, query policy, diagnostic resolution, scoring, and available evidence produced no additional calibrated exact commitments. The difficulty therefore goes beyond retrieving the correct diagnosis. Even when the diagnosis is present in the candidate set, the tested scorers often do not distinguish it strongly enough from similar diagnoses. Revision improves coverage at the broader category level, but it does not reliably recover the exact diagnosis.

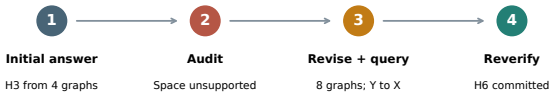

Figure 5: **An archived feedback example.** None of the initial four graphs is compatible. Within the protocol's single repair round, expansion and a probe returning effect $Y \to X = 0.003$ lead the LLM from H3 ($Y \to X$) to H6 ($X \to Y$), which passes reverification. Labels identify nodes and graphs in this illustrative controlled task; aggregate outcomes are in the CLadder protocol paragraph of Appendix C.

D.7 Calibrated comparators

At matched five-probe evidence without expansion, verification gives 23 of 120 correct commitments and one wrong, versus one correct and nine wrong for the 7B log-probability gate. Self-consistency commits none. These 23 correct commitments are distinct from the 23 retained answers in the expansion-only arm of Figure 2.

Every sealed run also froze calibrated one-score gates on the same scores, splits, and level. These include centroid and margin gates in the chambers; global-distance, cosine, fixed-multiscale, and distance-margin gates on the screens; log-probability and self-consistency on CLadder; and split conformal on every score. The kNN, Mahalanobis, and energy detectors were added post hoc. Table 19 displays the archived comparisons by setting, including the evidence budget and the distinction between single-gate and dual-gate arms.

Table 20 reports the comparisons paired with selection verification. At nine light-tunnel sensors, CAVEAT gives 17.2% correct commitment at 97% precision, compared with 8.0% at 76% for the margin gate with the same selection gate, and 15.7% at 98% for the Mahalanobis comparator. At this operating point, CAVEAT has the highest correct-commitment rate among these paired rules. On the 1,000-gene screens, the kNN, Mahalanobis, and energy detectors paired with selection verification each achieve less than half CAVEAT's correct-commitment rate.

On CLadder, the verifier alone with the same expansion and probe makes 69 correct and eight wrong commitments, including five wrong in 31 inadequate-space worlds, above the nominal level. Tightening its space gate to .05 gives three wrong in 30 such worlds. The reasoner's veto changes the original verifier-alone outcome to 42 correct and one wrong commitment.

Table 21 examines Proposition 3. In the light tunnel, the selection threshold borrowed from space nulls is at least the matched threshold in 28/29 rotations with three sensors and 29/29 with all sensors. It passes 42/184 and 51/332 compatible selections, compared with 163/184 and 308/332 under matched calibration; 123 and 257 units lie between the thresholds. Selection-null passes are 2/251 and 0/103, compared with 24/251 and 14/103 under matched calibration. The wind-tunnel comparison is less uniform. The borrowed threshold is at least the matched one in only 4/21 and 2/21 rotations; the table shows the corresponding compatible and null pass counts. Most wind-tunnel rotations have too few selection nulls for a finite threshold, which limits both-gate power.

Table 19: **Archived comparator results by setting.** Rates are percentages with archived 95% intervals. Wrong under omission uses the truth-absent/inadequate subgroup; wrong and correct over all units include noncommitments. Precision is conditional on commitment; n is the number committed. Rows distinguish evidence budgets and, in the physical and biological settings, space-gate-only (V) and dual-gate (V+S) rules. Physical and biological intervals follow the archived family procedures; clinical and LLM intervals are the archived binomial intervals.

Policy	Evidence	Wrong under omission	Wrong, all units	Correct, all units	Precision (%, n)
Causal Chamber					
Forced	3 sensors	100.0 [100.0, 100.0]	71.2 [67.1, 75.2]	28.8 [17.1, 41.4]	28.8 (638)
Always abstain	3 sensors	0.0 [0.0, 11.9]	0.0 [0.0, 11.9]	0.0 [0.0, 11.9]	–

Continued on next page

Table 19 continued

Policy	Evidence	Wrong under omission	Wrong, all units	Correct, all units	Precision (%, n)
Centroid gate	3 sensors	19.7 [7.9, 33.0]	14.6 [11.4, 17.6]	0.2 [0.0, 0.5]	1.1 (94)
Margin gate	3 sensors	9.9 [2.0, 20.2]	3.4 [1.6, 6.3]	6.6 [2.2, 11.8]	65.6 (64)
Centroid gate	9 sensors	14.3 [4.4, 26.6]	13.5 [10.0, 17.4]	1.6 [0.0, 4.1]	10.4 (96)
Margin gate	9 sensors	7.9 [1.5, 16.7]	2.5 [1.1, 3.9]	8.0 [3.1, 13.9]	76.1 (67)
kNN gate	3 sensors	16.3 [6.4, 27.6]	12.1 [9.1, 15.2]	4.1 [0.6, 8.8]	25.2 (103)
kNN gate	9 sensors	14.3 [3.9, 26.6]	12.7 [9.1, 16.8]	5.5 [2.2, 9.2]	30.2 (116)
Mahalanobis gate	3 sensors	4.4 [0.0, 11.3]	3.1 [1.4, 5.6]	1.9 [0.0, 5.5]	37.5 (32)
Mahalanobis gate	9 sensors	8.4 [1.0, 18.2]	6.1 [3.8, 8.8]	19.3 [11.8, 27.7]	75.9 (162)
Energy gate	3 sensors	8.9 [2.5, 16.7]	8.3 [5.2, 11.6]	0.3 [0.0, 0.9]	3.6 (55)
Energy gate	9 sensors	13.3 [4.9, 23.6]	9.6 [5.0, 14.6]	3.3 [0.3, 7.8]	25.6 (82)
Split conformal	3 sensors	17.7 [5.9, 31.5]	5.6 [4.2, 7.2]	18.3 [9.1, 28.5]	76.5 (153)
Split conformal	9 sensors	33.5 [18.7, 48.8]	10.7 [8.0, 13.5]	38.4 [27.7, 48.7]	78.3 (313)
CAVEAT (V only)	3 sensors	8.4 [2.0, 16.3]	10.0 [7.4, 12.7]	3.6 [0.0, 8.9]	26.4 (87)
CAVEAT (V only)	+1 selected sensor	8.4 [2.0, 16.3]	10.0 [7.4, 12.7]	3.6 [0.0, 9.4]	26.4 (87)
CAVEAT (V only)	9 sensors	10.3 [2.0, 21.2]	8.0 [4.5, 11.9]	19.6 [11.4, 28.7]	71.0 (176)
CAVEAT (V+S)	3 sensors	0.0 [0.0, 11.9]	0.5 [0.0, 1.1]	3.0 [0.0, 8.3]	86.4 (22)
CAVEAT (V+S)	9 sensors	1.5 [0.0, 3.4]	0.5 [0.0, 1.4]	17.2 [9.6, 26.0]	97.3 (113)
Wind tunnel					
Forced	3 sensors	100.0 [100.0, 100.0]	49.1 [46.1, 52.1]	50.9 [38.4, 62.2]	50.9 (336)
Always abstain	3 sensors	0.0 [0.0, 16.1]	0.0 [0.0, 16.1]	0.0 [0.0, 16.1]	–
Centroid gate	3 sensors	14.3 [5.7, 24.8]	11.3 [6.8, 16.7]	5.1 [1.2, 10.1]	30.9 (55)
Margin gate	3 sensors	10.5 [1.9, 21.0]	3.3 [0.9, 6.0]	3.9 [1.2, 6.8]	54.2 (24)
Centroid gate	11 sensors	20.0 [7.6, 35.2]	6.2 [3.0, 10.1]	17.9 [8.9, 28.3]	74.1 (81)
Margin gate	11 sensors	12.4 [2.9, 24.8]	3.9 [1.8, 6.2]	8.0 [3.6, 13.4]	67.5 (40)
kNN gate	3 sensors	17.1 [4.8, 31.4]	17.3 [12.8, 22.0]	17.3 [8.9, 26.5]	50.0 (116)
kNN gate	11 sensors	5.7 [0.0, 13.3]	5.7 [3.3, 8.3]	26.5 [15.5, 38.1]	82.4 (108)
Mahalanobis gate	3 sensors	16.2 [2.9, 31.4]	14.6 [10.4, 19.0]	14.9 [5.4, 25.9]	50.5 (99)
Mahalanobis gate	11 sensors	7.6 [1.0, 18.1]	3.9 [2.1, 5.7]	36.9 [25.3, 48.5]	90.5 (137)
Energy gate	3 sensors	16.2 [4.8, 29.5]	18.5 [13.7, 23.8]	13.7 [6.0, 22.6]	42.6 (108)

Continued on next page

Table 19 continued

Policy	Evidence	Wrong under omission	Wrong, all units	Correct, all units	Precision (%, n)
Energy gate	11 sensors	6.7 [1.9, 13.3]	2.4 [0.6, 4.5]	27.7 [19.0, 36.9]	92.1 (101)
Split conformal	3 sensors	25.7 [11.4, 41.9]	8.0 [6.0, 10.1]	24.7 [13.7, 36.0]	75.5 (110)
Split conformal	11 sensors	22.9 [9.5, 39.0]	7.1 [5.4, 8.9]	50.0 [39.6, 59.5]	87.5 (192)
CAVEAT (V only)	3 sensors	13.3 [1.0, 26.7]	16.4 [11.3, 21.4]	18.2 [8.0, 29.2]	52.6 (116)
CAVEAT (V only)	11 sensors	7.6 [1.0, 18.1]	2.7 [0.9, 4.8]	40.5 [28.3, 52.4]	93.8 (145)
CAVEAT (V+S)	3 sensors	1.9 [0.0, 5.7]	0.6 [0.0, 1.5]	3.3 [0.9, 5.7]	84.6 (13)
CAVEAT (V+S)	11 sensors	1.0 [0.0, 2.9]	0.3 [0.0, 0.9]	2.1 [0.9, 3.3]	87.5 (8)
Jurkat CRISPRi					
Forced	1,000 genes	100.0 [100.0, 100.0]	83.9 [80.7, 86.8]	16.1 [13.1, 19.2]	16.1 (720)
Always abstain	1,000 genes	0.0 [0.0, 2.0]	0.0 [0.0, 2.0]	0.0 [0.0, 2.0]	–
Global-distance gate	1,000 genes	10.0 [6.2, 14.0]	8.5 [5.1, 11.9]	3.6 [1.9, 5.6]	29.9 (87)
Cosine gate	1,000 genes	8.6 [5.1, 12.4]	6.5 [3.9, 9.4]	5.4 [3.3, 7.6]	45.3 (86)
Fixed-multiscale gate	1,000 genes	6.0 [3.1, 9.2]	4.6 [2.4, 7.2]	5.6 [3.5, 7.6]	54.8 (73)
Distance-margin gate	1,000 genes	7.6 [4.6, 11.2]	5.4 [3.2, 7.9]	6.9 [4.7, 9.4]	56.2 (89)
kNN gate	1,000 genes	10.2 [6.4, 14.4]	9.4 [6.0, 13.2]	2.8 [1.2, 4.4]	22.7 (88)
Mahalanobis gate	1,000 genes	7.6 [4.4, 11.2]	7.5 [4.3, 11.1]	2.4 [1.0, 3.9]	23.9 (71)
Energy gate	1,000 genes	10.0 [6.2, 14.1]	8.3 [5.0, 11.8]	3.1 [1.5, 4.9]	26.8 (82)
Split conformal	1,000 genes	4.0 [2.1, 6.4]	2.9 [1.4, 4.6]	2.1 [1.0, 3.3]	41.7 (36)
CAVEAT (V only)	1,000 genes	8.6 [5.1, 12.3]	7.5 [4.4, 10.8]	5.7 [3.6, 7.9]	43.2 (95)
CAVEAT (V+S)	1,000 genes	1.4 [0.2, 2.9]	1.1 [0.3, 2.2]	3.1 [1.5, 4.7]	73.3 (30)
HepG2 CRISPRi					
Forced	1,000 genes	100.0 [100.0, 100.0]	75.3 [71.7, 78.9]	24.7 [21.1, 28.3]	24.7 (360)
Always abstain	1,000 genes	0.0 [0.0, 2.0]	0.0 [0.0, 2.0]	0.0 [0.0, 2.0]	–
Global-distance gate	1,000 genes	6.8 [3.6, 10.5]	6.7 [3.3, 10.3]	1.9 [0.6, 3.3]	22.6 (31)
Cosine gate	1,000 genes	8.9 [5.2, 13.4]	6.7 [3.6, 10.0]	8.1 [5.3, 10.8]	54.7 (53)
Fixed-multiscale gate	1,000 genes	6.3 [3.1, 10.0]	5.8 [2.8, 9.4]	6.9 [4.4, 9.4]	54.3 (46)
Distance-margin gate	1,000 genes	7.9 [4.3, 12.0]	4.4 [2.5, 6.9]	9.4 [6.7, 12.5]	68.0 (50)
kNN gate	1,000 genes	7.9 [4.2, 11.8]	8.1 [4.4, 11.9]	1.1 [0.3, 2.2]	12.1 (33)
Mahalanobis gate	1,000 genes	7.4 [3.7, 11.2]	6.4 [3.3, 10.0]	3.9 [1.9, 5.8]	37.8 (37)
Energy gate	1,000 genes	6.8 [3.6, 10.6]	6.7 [3.3, 10.3]	0.8 [0.0, 1.9]	11.1 (27)

Continued on next page

Table 19 continued

Policy	Evidence	Wrong under omission	Wrong, all units	Correct, all units	Precision (%, n)
Split conformal	1,000 genes	8.4 [4.7, 12.7]	5.0 [2.8, 7.8]	4.7 [2.8, 6.9]	48.6 (35)
CAVEAT (V only)	1,000 genes	7.4 [4.0, 11.2]	6.1 [3.1, 9.4]	8.9 [6.1, 11.7]	59.3 (54)
CAVEAT (V+S)	1,000 genes	2.1 [0.5, 4.3]	1.7 [0.3, 3.6]	7.5 [5.0, 10.3]	81.8 (33)
LiveClin					
Forced	text + up to 3 stems	100.0 [97.4, 100.0]	93.6 [88.8, 96.7]	6.4 [3.3, 11.2]	6.4 (171)
Always abstain	text + up to 3 stems	0.0 [0.0, 2.7]	0.0 [0.0, 2.1]	0.0 [0.0, 2.1]	–
One-class gate	text + up to 3 stems	5.0 [2.0, 10.0]	4.7 [2.0, 9.0]	0.6 [0.0, 3.2]	11.1 (9)
Margin gate	text + up to 3 stems	19.1 [13.0, 26.6]	16.4 [11.2, 22.8]	2.3 [0.6, 5.9]	12.5 (32)
Entropy gate	text + up to 3 stems	19.1 [13.0, 26.6]	16.4 [11.2, 22.8]	2.9 [1.0, 6.7]	15.2 (33)
Ordinary conformal	text + up to 3 stems	0.0 [0.0, 2.6]	0.0 [0.0, 2.1]	1.2 [0.1, 4.2]	100.0 (2)
CAVEAT	text + up to 3 stems	0.0 [0.0, 2.7]	0.0 [0.0, 2.1]	0.6 [0.0, 3.2]	100.0 (1)
CLadder					
Forced	5 probes	100.0 [94.0, 100.0]	75.8 [67.2, 83.2]	24.2 [16.8, 32.8]	24.2 (120)
Always abstain	5 probes	0.0 [0.0, 6.0]	0.0 [0.0, 3.0]	0.0 [0.0, 3.0]	–
Vocabulary gate, no expansion	5 probes	6.7 [1.8, 16.2]	15.8 [9.8, 23.6]	20.8 [14.0, 29.2]	56.8 (44)
Dual gate, no expansion	5 probes	1.7 [0.0, 8.9]	0.8 [0.0, 4.6]	19.2 [12.6, 27.4]	95.8 (24)
Reasoner may answer NONE	5 probes	98.3 [91.1, 100.0]	77.5 [69.0, 84.6]	20.0 [13.3, 28.3]	20.5 (117)
NONE + calibrated confidence gate	5 probes	0.0 [0.0, 6.0]	0.0 [0.0, 3.0]	0.0 [0.0, 3.0]	–
Log-probability gate (Qwen2.5-7B)	5 probes	6.7 [1.8, 16.2]	7.5 [3.5, 13.8]	0.8 [0.0, 4.6]	10.0 (10)
Self-consistency gate (Qwen2.5-7B)	5 probes	0.0 [0.0, 6.0]	0.0 [0.0, 3.0]	0.0 [0.0, 3.0]	–
Log-probability gate (Qwen2.5-32B)	5 probes	10.0 [3.8, 20.5]	7.5 [3.5, 13.8]	2.5 [0.5, 7.1]	25.0 (12)
Self-consistency gate (Qwen2.5-32B)	5 probes	0.0 [0.0, 6.0]	0.0 [0.0, 3.0]	0.0 [0.0, 3.0]	–
CAVEAT	5 probes + query	3.2 [0.1, 16.7]	0.8 [0.0, 4.6]	35.0 [26.5, 44.2]	97.7 (43)
Verifier alone, expansion + probe	5 probes + query	16.1 [5.5, 33.7]	6.7 [2.9, 12.7]	57.5 [48.1, 66.5]	89.6 (77)

Table 20: **Comparators paired with selection verification.** Nine-sensor light-tunnel and 1,000-gene screen comparisons cited in the text. Each row combines its space/detector rule with the archived selection check. These differ from the gate-alone rows in Table 19. Entries are percentages with archived 95% intervals; precision is conditional on commitment and n counts commitments.

Space/detector rule	Wrong under omission	Wrong, all units	Correct, all units	Precision (%, n)
Causal Chamber (9 sensors)				
CAVEAT	1.5 [0.0, 3.4]	0.5 [0.0, 1.4]	17.2 [9.6, 26.0]	97 (113)
Margin gate	7.9 [1.5, 16.3]	2.5 [1.1, 3.9]	8.0 [3.1, 13.9]	76 (67)
Centroid gate	8.9 [1.0, 19.7]	4.7 [3.4, 6.0]	0.3 [0.0, 0.9]	6 (32)
kNN gate	3.4 [1.0, 6.9]	2.7 [1.4, 4.1]	2.7 [0.9, 4.7]	50 (34)
Mahalanobis gate	1.0 [0.0, 2.5]	0.3 [0.0, 0.9]	15.7 [8.9, 23.0]	98 (102)
Energy gate	3.9 [1.0, 7.9]	1.7 [0.3, 3.4]	1.1 [0.0, 2.5]	39 (18)
Jurkat CRISPRi (1,000 genes)				
CAVEAT	1.4 [0.2, 2.9]	1.1 [0.3, 2.2]	3.1 [1.5, 4.7]	73 (30)
Global-distance gate	1.0 [0.2, 2.0]	0.8 [0.1, 1.8]	1.5 [0.4, 2.8]	65 (17)
Cosine gate	4.0 [1.9, 6.7]	2.4 [1.1, 3.9]	3.1 [1.5, 5.0]	56 (39)
Fixed-multiscale gate	1.0 [0.0, 2.2]	0.7 [0.0, 1.7]	3.1 [1.7, 4.7]	81 (27)
Distance-margin gate	7.6 [4.6, 11.2]	5.4 [3.2, 7.9]	6.9 [4.7, 9.3]	56 (89)
kNN gate	0.2 [0.0, 0.7]	0.3 [0.0, 0.8]	1.1 [0.3, 2.2]	80 (10)
Mahalanobis gate	0.5 [0.0, 1.2]	0.4 [0.0, 1.1]	0.6 [0.0, 1.2]	57 (7)
Energy gate	0.7 [0.0, 1.7]	0.7 [0.0, 1.7]	0.8 [0.1, 1.8]	55 (11)
HepG2 CRISPRi (1,000 genes)				
CAVEAT	2.1 [0.5, 4.3]	1.7 [0.3, 3.6]	7.5 [5.0, 10.3]	82 (33)
Global-distance gate	1.1 [0.0, 2.7]	0.6 [0.0, 1.4]	1.4 [0.3, 2.8]	71 (7)
Cosine gate	1.6 [0.0, 3.7]	0.8 [0.0, 1.9]	3.9 [1.9, 5.8]	82 (17)
Fixed-multiscale gate	3.2 [1.0, 5.8]	1.9 [0.6, 3.6]	6.4 [4.2, 8.9]	77 (30)
Distance-margin gate	7.9 [4.3, 12.0]	4.4 [2.5, 6.9]	9.4 [6.7, 12.5]	68 (50)
kNN gate	1.1 [0.0, 2.7]	0.8 [0.0, 2.2]	0.6 [0.0, 1.4]	40 (5)
Mahalanobis gate	0.5 [0.0, 1.6]	0.3 [0.0, 0.8]	2.5 [1.1, 4.2]	90 (10)
Energy gate	0.5 [0.0, 1.6]	0.3 [0.0, 0.8]	0.3 [0.0, 0.8]	50 (2)

D.8 THE REASONER PROPOSES THE SPACE

In the post-hoc Phase 21 study, the reasoner receives the story, named variables, and five base probes for the same 120 test worlds. It proposes up to four DAGs and selects one, without a supplied can-

Table 21: **Matched and borrowed selection thresholds.** Matched calibrates the selection statistic on selection failures; borrowed calibrates it on space failures. The archived statistic is a higher-is-better margin, passing when it exceeds its threshold. The threshold-order column counts rotations with a finite matched threshold and $\tau_{S|V} \geq \tau_{S|S}$. Compatible and selection-null pass columns apply the selection threshold alone, not the full commitment rule. Between counts compatible units whose margins lie between the two cutoffs, irrespective of their order.

Setting	Threshold order	Compatible matched	Compatible borrowed	Between	S-null matched	S-null borrowed
Light, 3 sensors	28/29	163/184	42/184	123	24/251	2/251
Light, 9 sensors	29/29	308/332	51/332	257	14/103	0/103
Wind, 3 sensors	4/21	28/171	13/171	23	3/60	0/60
Wind, 11 sensors	2/21	14/222	27/222	11	0/9	0/9

didate space. CAVEAT verifies the proposals with its frozen map. A space null has no compatible graph; a selection null has one but the reasoner selects an incompatible graph. On Reject/Revise, the certificate is returned and the reasoner proposes again. On Query, the CAVEAT probe is acquired. Every arm is reverified. Proposed edges use the same deterministic coefficients as supplied graphs, so both are scored identically.

We run Qwen2.5-7B, -14B, -32B, and phi-4, each in 4-bit precision. When an arm has fewer than three selection-null calibration worlds, the implementation reports V-only results. Small-null relaxed selection levels are .125 for 7B initially, $1/6$ after its revision, and .25 after phi-4 revision. These do not certify selection failure at .10. Table 22 labels the gate configuration for each reported arm.

Forced commitment is wrong in 70, 40, 31, and 32 of 120 worlds for 7B, 14B, 32B, and phi-4, respectively. Every commitment from an inadequate proposed space is wrong. Verification without revision reduces wrong commitments to five for 7B and zero for the other three models, using the gate configurations shown in the table. Certificate-guided revision raises proposal adequacy from 48% to 62% for 7B, 79% to 92% for 14B, and 78% to 81% for phi-4. The table reports correct commitments, inadequate-space errors, and precision for revision alone and with a random or selected probe; precision in these repair arms is 92–100%.

Table 22: **Verification when the reasoner proposes the candidates (post hoc).** Each model uses 120 test worlds. Adequacy is the percentage with a task-compatible proposal. Wrong/correct and exact-graph commitments are counts; wrong on inadequate spaces is a subgroup count/denominator. Precision is conditional on commitment, with the number committed in parentheses. Gate labels identify V-only versus V+S outputs; the relaxed selection levels are specified in the accompanying text. The identical repeated revision rows for Qwen2.5-32B are omitted.

Policy	Gate	Adeq. (%)	Wrong	Correct	Wrong on inadequate	Precision (%, n)	Exact
Qwen2.5-7B							
Forced	–	48	70	50	62/62	42 (120)	28
Verify, no revision	V+S	48	5	29	5/62	85 (34)	28
Revise	V+S	62	3	33	3/46	92 (36)	30
Revise + random probe	V+S	62	2	33	2/46	94 (35)	30
Revise + selected probe	V+S	62	2	33	2/46	94 (35)	30
Qwen2.5-14B							
Forced	–	79	40	80	25/25	67 (120)	64
Verify, no revision	V+S	79	0	64	0/25	100 (64)	64
Revise	V+S	92	0	75	0/10	100 (75)	67
Revise + random probe	V+S	92	0	79	0/10	100 (79)	68
Revise + selected probe	V+S	92	0	76	0/10	100 (76)	68

Continued on next page

Table 22 continued

Policy	Gate	Adeq. (%)	Wrong	Correct	Wrong on inadequate	Precision (%, n)	Exact
Qwen2.5-32B							
Forced	–	75	31	89	30/30	74 (120)	71
Verify, no revision	V	75	0	75	0/30	100 (75)	71
phi-4							
Forced	–	78	32	88	26/26	73 (120)	40
Verify, no revision	V	78	0	42	0/26	100 (42)	40
Revise	V+S	81	0	64	0/23	100 (64)	49
Revise + random probe	V+S	81	0	61	0/23	100 (61)	50
Revise + selected probe	V+S	81	0	54	0/23	100 (54)	48

D.9 LIMITATIONS

The physical experiments have few independent families, 29 in the light tunnel and 21 in the wind tunnel, so their family-bootstrap intervals are wide, and wind-tunnel identification may partly reflect recording background rather than the causal signal (Appendix D.5). LiveClin is retrospective: published case reports rather than live cases, with the recorded diagnosis as the reference. The LLM trial controls candidate availability and ambiguity by construction, so its regimes are stress panels rather than a deployment mixture. The learned score has been validated in simulation only. The action guard is deliberately worst-case over the retained candidates; a weighted rule would raise coverage at the price of calibrated candidate probabilities, which the proposer does not supply. The null-specific guarantee is population-level and may fail under distribution shift without additional assumptions and recalibration (Barber et al., 2023); any change to the proposer, retrieval, score, query rule, or stopping rule defines a new pipeline and requires fresh calibration (Appendix B.4).