DO RESEARCH AGENTS IMPROVE FOR THE RIGHT REASONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

An autonomous research agent can improve a system through an accurate diagnosis, feedback-driven search, or a useful intervention based on an inaccurate account. Final performance alone cannot distinguish these possibilities. We introduce a controlled evaluation framework in simulated quantum error correction that separately tests the accuracy of an agent's external mechanism record and its contribution to neural-decoder improvement. The agent expresses a noise hypothesis as a typed physical record, which a fixed compiler turns into training data. We intervene on record content, vary access to performance feedback, and test transfer and revision on independently initialized target decoders. Accurate records provide incremental training value without target performance feedback, and revising a record after a mechanism change improves both diagnostic accuracy and downstream utility. Yet inaccurate records and diagnosis-free policies also improve performance, while a simple physical estimator remains competitive with the agent. These findings show why improvement is not evidence of correct diagnosis, and demonstrate how externally checkable reasons can be tested for their contribution to improvement. The resulting claim concerns the operational value of submitted mechanism representations, not an agent's private understanding or a general ability to conduct open-ended research.

1 INTRODUCTION

When a research agent makes a model better, what has it learned? A lower error rate establishes that an intervention worked. It does not establish that the agent correctly diagnosed the original limitation, that its stated diagnosis influenced the intervention, or that the diagnosis will remain useful when the environment changes. These distinctions matter if autonomous researchers are to do more than repeatedly try plausible edits and retain the winners.

Existing research-agent evaluations offer substantive tests of scientific workflows and iterative experimentation (Chen et al., 2025; Ning et al., 2026). Mechanism-oriented evaluations already go beyond final scores: CausalGame tests causal thinking in controlled environments, and BoxingGym evaluates experimental design and the predictive usefulness of communicated explanations (Chen et al., 2026; Gandhi et al., 2025). The question we study is narrower than whether agents can explain an outcome: *does an independently checkable diagnosis contribute to improving an external system?* Answering it requires measuring diagnosis and utility separately, and intervening on the link between them.

We construct such a test in simulated quantum error correction (QEC). A neural decoder predicts the logical effect of physical noise from observed error-detection signals. A language agent receives evidence about the noise and proposes a training-data intervention. Crucially, we know the physical generator and can change its correlation structure while keeping individual-qubit error rates fixed. The agent's mechanism estimate can therefore be evaluated against a target independent of whether its proposed training succeeds. Its proposal also has a real downstream consequence: training a separate neural network and measuring errors on fresh samples.

The central interface is an *executable mechanism record*. It contains physical parameters rather than unrestricted prose. A fixed compiler converts the submitted parameters into a data-generating program; the agent cannot edit the decoder, optimizer, training budget, or evaluator. This lets us replace a record, remove it, or require its revision while holding execution fixed. It does not make

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.

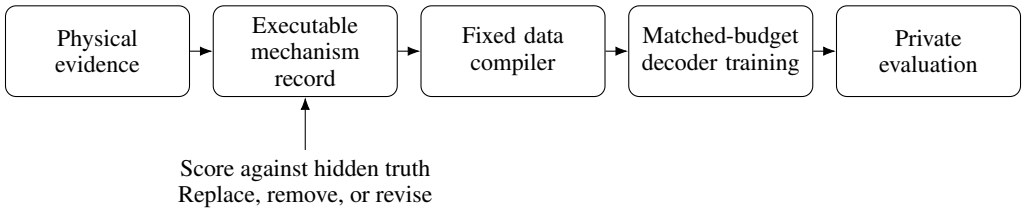


Figure 1: The central composition study. The final record is the only agent-controlled input to data synthesis. Its accuracy and downstream value are measured separately. All target proposals precede target training; no target performance is returned to the agent. Earlier decision-interface studies also expose evidence to the downstream action selector and are not strict record-only interfaces.

a correct record necessary for improvement. Testing that distinction is a principal purpose of the framework.

Our completed studies yield a qualified positive answer. In a composition-and-transfer study with independently initialized target models, accurate source records help even when no target performance feedback is available. Perturbing their physical content reduces their value. When new observations reveal a changed mechanism, updating the record improves both its accuracy and the resulting training intervention. However, inaccurate records and ordinary training also improve average performance. A separate feedback-budget experiment shows that search itself is useful. Finally, the agent's behavior is largely compatible with simple physical estimation and mapping rules; we do not establish an LLM-specific advantage.

The contribution is a measurement framework and an empirical separation, not a claim to observe internal understanding. We distinguish questions that an endpoint score conflates: whether a diagnosis is accurate, whether its content changes the intervention, and whether that change improves the external system. We then test whether this relationship survives transfer and revision. Positive and negative results jointly constrain the answer: external reasons can have incremental value, but neither improvement nor a diagnosis-to-action interface certifies that the agent improved *because it understood*.

2 What would count as improving for the right reasons?

We use *reason* to mean a submitted, falsifiable representation of an environmental mechanism. This differs from a private belief or a complete account of internal computation. An agent may produce the same record by reasoning, pattern matching, or executing a familiar statistical rule. Our experiments identify properties of observable research behavior, not which internal process produced it.

Three measurements are required. First, *accuracy*: does the record agree with a hidden, independently specified generator? Second, *content dependence*: does changing its physical content change intervention quality, beyond a meaning-preserving presentation change? Third, *utility*: does the intervention improve a held-out endpoint under matched budgets? Transfer and revision provide further tests: a useful mechanism representation should support appropriate action when geometry changes, and should be revisable when new evidence contradicts its parameters.

None of these tests subsumes the others. An accurate record can compile into a poor data policy. An inaccurate record can produce useful examples. An agent can ignore a correct record, or follow an incorrect one faithfully. A strong endpoint can also be reached by feedback search. We avoid assigning individual runs psychological labels such as "understands" or "lucky." Instead, we report controlled differences between observable interfaces.

Why an error trace is not enough. A simulator records what happened, but a realized fault is not automatically a learnable decoder defect. Different physical errors can produce the same syndrome, and a logical failure depends on both the physical realization and the decoder's response. Removing a fault and replaying the same decoder establishes dependence under that deletion; it does not establish a unique root cause or show that more examples of the fault improve native-distribution

performance. Our exact-distribution audit makes this concrete: even a Bayes-optimal decoder has failures that physical deletion rescues (Appendix C). The central study therefore evaluates estimates of known generating parameters, not purported unique causes of individual failures.

3 A CONTROLLED RESEARCH ENVIRONMENT

3.1 WHY QUANTUM ERROR CORRECTION?

QEC supplies a useful combination of observable signals, hidden physical noise, and a consequential learning problem. The decoder sees a syndrome—a pattern of detection events—and predicts whether the logical observable has flipped. Stabilizer simulation permits efficient sampling of these quantities (Gidney, 2021). Neural decoding is an established application of machine learning (Bausch et al., 2024); here it is the external system the agent tries to improve, not a claim of a new state-of-the-art decoder.

Our particular advantage is experimental control. We can vary joint physical errors without changing each qubit's marginal error probability, move the same parameters to a different geometry, and retain the generating parameters for hidden scoring. Thus, "the system has more errors" cannot explain every task difference. The mapping from noise to syndrome and logical error still goes through an actual simulated correction circuit, and a data policy must earn its utility through neural-network training.

This combination makes QEC a tractable substrate for the measurement question. We do not claim that ground-truth mechanisms are unique to quantum systems, nor that simulation establishes performance on hardware. The central experiments use a small surface-code memory task and a declared noise family, permitting controlled comparisons at manageable computational cost.

3.2 MECHANISM RECORDS AND INTERVENTION COMPILATION

The central study places local pair-noise channels at known circuit locations. A coefficient controls how strongly the two qubits in a pair fail together, while their individual error probabilities remain fixed. A record contains the estimated coefficient for each of three physical pairs. Four supported geometries differ in pair placement. The task is a bounded estimation-and-transfer problem, not discovery of an unrestricted physical law.

The source agent observes physical-channel outcome counts and submits its estimate through a typed schema. These counts are privileged simulator evidence, not measurements available from syndrome data alone. A target agent receives a source record or a controlled replacement, an explicit geometry mapping, and, in some conditions, fresh target counts. It returns a final physical data program. Estimates are bounded integers with a declared conversion to physical coefficients; explanatory prose cannot alter execution.

The compiler receives only that final program and fixed public circuit information. It does not read the evidence, record condition, hidden generator parameters, previous outcomes, or task identity. All policies share a fixed synthesis rule mixing ordinary samples with syndrome-hard samples from the submitted model. The rule, decoder architecture, optimizer settings, and update budget are fixed before evaluation. Decoder weights do change during training; "fixed training stack" does not mean frozen decoder weights. The hosted agent uses a fixed model alias and configuration, not a verifiably frozen service-side weight snapshot.

3.3 TWO DISTINCT OUTCOMES

Diagnostic error is the mean squared difference between submitted and generating coefficients. Training utility is the log ratio of initial to final logical error rate (LER):

$$D(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) = \frac{1}{3}\sum_{j=1}^{3}(\widehat{\alpha}_j - \alpha_j)^2, \qquad U = \log \frac{\widetilde{\mathrm{LER}}_{\mathrm{initial}}}{\widetilde{\mathrm{LER}}_{\mathrm{trained}}}. \tag{1}$$

The tilde denotes fixed small-count smoothing (Appendix A). Lower D and higher U are better. Positive utility means improvement over initialization; a positive *between-policy difference* could

instead mean less harm. We report both. Neither metric is defined using the other, so their disagreement is informative rather than impossible by construction.

4 EXPERIMENTAL DESIGN

4.1 THE CENTRAL TEST: TRANSFER WITHOUT TARGET PERFORMANCE FEEDBACK

The composition study uses 48 independently initialized target decoders across a fixed grid of geometries and held-out coefficient compositions. All source responses and target proposals are committed before target training or private evaluation. The agent cannot choose the reported intervention by observing target training success. This removes target-end performance trial-and-error, not internal computation or the research team's earlier development work.

We compare three source-record conditions: the original estimate, a cyclic permutation of its coefficients, and no source record. A permutation is not declared incorrect by fiat: its diagnostic error is measured against the hidden generator. We cross these conditions with three target regimes:

- *Preserved mechanism, no target observations.* Preservation under a known geometry change is guaranteed. A record may be transferred, but cannot be re-estimated from target counts.
- *Preserved mechanism, target observations.* New counts allow confirmation or refinement of the transferred record.
- *Changed mechanism, target observations.* New counts allow revision of stale coefficients. The prompt does not reveal a private change label.

The clean content comparison is original versus permuted record within the first regime. The revision comparison is a newly estimated program versus the correctly mapped but unrevised source record within the changed regime. Comparing the first and third regimes would change both evidence and environment, and is not treated as an isolated evidence effect.

Controls. We include deterministic source-record mapping, literal source-program reuse, a non-LLM count estimator, and a diagnostic oracle supplied with true generating coefficients. Additional controls use homogeneous coefficients, a broad mechanism mixture, and native-distribution training. The diagnostic oracle is not an oracle for the best finite-budget training policy. The count estimator is a whole-pipeline reference with access to the corresponding observations; it is not an equally informed recipient for the no-record/no-target-observations condition.

Each intervention starts from the same target checkpoint and optimizer state within a block. Training seeds and fresh evaluation panels are paired across policies. Three training repeats are averaged within each block before inference. Exactly identical programs are executed once per block and repeat, then share the fit; they are not independent replications. The study contains 1,983 distinct fits, but its inferential sample remains 48 blocks.

Analysis discipline. Before formal execution, we fixed six primary contrasts: source accuracy, record value, record-content value, diagnostic revision, useful revision, and advantage over the homogeneous policy. They form one Holm-corrected family. Registrations are timestamped local manifests and hashed source/input sets, not externally certified public preregistrations. We retain all blocks and operational failures under declared policies. Appendix A specifies coverage requirements and finite-grid inference limits; Appendix B reports all control means.

4.2 SUPPORTING TESTS: INFORMATION, SEARCH, AND CHANGED WORLDS

A separate feedback-budget study crosses outcome-only versus richer counterfactual evidence with fixed opportunities to query trial outcomes. It asks whether feedback search helps without rich diagnosis evidence, and whether rich evidence helps before such search. This study selects interventions from frozen execution banks rather than claiming unrestricted code-writing autonomy. Its action interface can also see evidence, so an evidence effect is not automatically mediated entirely by the record.

Table 1: Central study: all six prespecified primary contrasts. Positive values favor the stated hypothesis. Intervals are pointwise 95% block-level intervals; p values are Holm-adjusted within this family. MSE reductions and utility differences have different units and are not percentages. Appendix B gives simultaneous intervals.

Contrast	Effect	95% interval	Adjusted p
Source accuracy over uninformative estimate (MSE)	.1044	[.1037, .1052]	$< .001$
Original versus no record, no target observations (U)	.0060	[.0031, .0088]	$< .001$
Original versus permuted record, same regime (U)	.0275	[.0199, .0352]	$< .001$
Revision versus stale mapping, changed world (MSE)	.4240	[.4144, .4336]	$< .001$
Revision versus stale mapping, changed world (U)	.0386	[.0298, .0474]	$< .001$
Original versus homogeneous, regime average (U)	.0058	[.0042, .0073]	$< .001$

Additional completed tests intervene on record content at a fixed decision interface, assess cross-task memory transfer, and require prospective predictions under physical world changes. Meaning-preserving serialization or label-renaming controls distinguish semantic changes from presentation effects. We retain these studies separately because they differ in information access, target reuse, and inferential units. They are complementary evidence, not samples to pool into one larger experiment.

5 Results

5.1 Improvement does not certify diagnostic correctness

Without target observations, both original and permuted source records produce positive average training utility. Yet the permuted records are substantially farther from the true coefficients, even relative to an uninformative constant estimate. Native training and the no-record policy also improve the model. The combination "inaccurate external diagnosis, improved external system" occurs in the same controlled evaluation that rewards accurate records.

This does not assert that the agent privately believed a false explanation. The perturbed record is an experimental input, and its accuracy is an external property. The result establishes that performance improvement cannot validate that record. Conversely, accurate records have greater utility on average, so the dissociation does not imply that diagnosis is irrelevant.

5.2 Record content has incremental value without target search

All six central primary contrasts pass the prespecified correction and coverage gates (Table 1). Without target observations, the original record outperforms both its permuted version and the no-record condition. This separates the value of useful information from the stronger disruption caused by inaccurate information. The original-versus-no-record effect is modest, much smaller than the original policy's total improvement over initialization.

Because proposals precede target training, these differences cannot be explained solely by selecting winners from target performance feedback. Paired initial states, budgets, and evaluation samples also prevent extra training or a selectively favorable evaluation draw from defining the comparison. Nevertheless, this identifies the average effect of the record condition in this interface. It neither assigns the whole improvement to diagnosis nor measures a fraction mediated by private understanding.

5.3 Revision improves both the record and the intervention

When target counts reveal changed coefficients, the agent revises its record toward the new generator. The revised program produces better training utility than a geometry-correct but stale program. Both parts matter: a better estimate without a better decoder establishes diagnostic revision only; a better decoder without a better estimate does not establish accuracy-linked revision.

The original-record agent improves over initialization in every regime on average. Its advantages are not solely differences between harmful policies. However, a source record does not always

Table 2: Evidence and search study: task-level private log-LER gain contrasts over 24 tasks. These rows form their own Holm family. Evidence is outcome-only (E_0) or counterfactual (E_3); B is feedback-query budget.

Contrast	Effect	95% interval	Adjusted p
Search: $(B_8 - B_0)$ under E_0	.0306	[.0129, .0514]	.0115
Evidence: $(E_3 - E_0)$ at B_0	.0869	[.0566, .1221]	$< .001$
Interaction: $(E_3 - E_0)_{B_8} - (E_3 - E_0)_{B_0}$	-.0302	[-.0507, -.0120]	.0115

help when new evidence is available. In the changed regime, the no-record agent has slightly higher descriptive mean utility than the original-record agent. The conclusion is that revision beats stubborn reuse, not that memory always helps or that revising old knowledge beats estimating afresh.

5.4 Search and evidence both contribute

The feedback-budget experiment gives outcome-only agents a measurable benefit from additional trial feedback. Richer evidence also improves zero-feedback decisions. The interaction is negative: the rich-evidence advantage becomes smaller at the larger search budget (Table 2). This is compatible with partial substitution between informative evidence and feedback search within the tested interface. It does not imply that feedback is inherently harmful or that all evidence benefits flow through a diagnosis record.

5.5 Simple physical rules explain much of the successful behavior

The non-LLM count estimator is competitive with the agent, with slightly higher mean utility in the central study. Under guaranteed preservation without target observations, the original-record program agrees exactly with deterministic geometry mapping in every block. Equal utility reflects exact-program deduplication, not two independent experiments demonstrating equivalence. An earlier changed-world study likewise shows high descriptive agreement with a simple physical routing rule.

We have not demonstrated an advantage requiring language-model reasoning, unknown-mechanism discovery, or invention of a training algorithm. At the same time, a physically correct rule can itself use the right information. Agreement does not invalidate that information's measured value. The contribution is making this distinction visible rather than interpreting a positive agent score or a successful record intervention as uniquely sophisticated reasoning.

5.6 The evidence does not support a uniformly positive story

An earlier cross-task transfer study did not establish any of its three primary memory-transfer advantages after correction. Those null results remain: the later composition study uses a different, explicitly mapped design and independent initialization, and does not retroactively validate the earlier claim. An initial changed-world experiment also suffered service-related coverage failures. A complete service replication supports prospective prediction and renaming equivalence, but uses the same physical task panel and archived models. It is not independent-task replication and does not repair the original denominators.

The earliest predicate-based closed-loop study failed to confirm its three positive primary hypotheses. Its strong ordinary-data baseline and subsequent exact-distribution audit motivated separating physical attribution from learnable training value. These findings explain why our claim is narrower than "richer evidence causes understanding, which causes improvement." Primary outcomes and failure handling are retained in Appendix C.

6 Related work

Research agents and scientific discovery. ScienceAgentBench evaluates scientific workflow tasks with executable outputs and external evaluation (Chen et al., 2025). Specialist-agent autore-

search records hypotheses, code edits, lineage, and evaluator feedback in closed empirical loops (Ning et al., 2026). These systems study broader action spaces. Our setting trades breadth for the ability to score physical estimates independently and intervene on their contribution to downstream training. We do not characterize all prior work as measuring only a final scalar.

Mechanism evaluation and explanation transfer. CausalGame separates successful solutions from causal thinking under controlled generative mechanisms (Chen et al., 2026). BoxingGym evaluates experiment selection, model discovery, and whether explanations support another agent's predictions (Gandhi et al., 2025). These are direct precedents for evaluating more than outcomes and for giving explanations a functional role. Our complementary endpoint is whether a manipulated mechanism record helps train an independent neural system, including transfer without target outcome feedback and useful revision after physical change.

Intervenable representations and faithfulness. Concept bottleneck models expose intermediate concepts that can be edited to alter predictions (Koh et al., 2020). Chain-of-thought faithfulness studies intervene on reasoning text, including errors and paraphrases, to test downstream dependence (Lanham et al., 2023). An editable representation or an error-versus-placebo comparison is not independently novel. Here the record has evaluator-known physical semantics and is compiled into an external learning intervention. Dependence on this external record still does not identify internal computation.

Quantum simulation and neural decoding. Stim enables efficient stabilizer simulation (Gidney, 2021); AlphaQubit demonstrates the practical relevance of learned decoding (Bausch et al., 2024). We reuse this application as a controlled research task. Our small decoder and synthetic noise family do not reproduce hardware results or establish a decoding advance. The mechanism-to-training pathway, rather than quantum performance, is the object of study.

7 Discussion and limitations

The right-reasons question is most useful when decomposed into independently falsifiable claims. We can establish that a submitted estimate is accurate, that its content has incremental value, and that evidence-driven revision improves an intervention. We cannot infer that every success used a correct diagnosis, that all improvement comes from diagnosis, or that manipulating a record randomizes private understanding. Dividing a record contrast by total utility would not yield a cognitive attribution percentage.

The environment is constrained. The noise family, parameter meanings, coordinate mapping, compiler, and training stack are supplied. The agent is one hosted model configuration with a tool-free decision interface. It neither chooses an arbitrary scientific question nor writes unrestricted experimental code. Privileged physical observations are stronger than syndrome-only access, and the central task is estimation within a known family. Results do not establish behavior across frontier agents, code sizes, real-device noise, or open-ended research.

The fixed compiler permits controlled content interventions but embeds human-designed knowledge about converting mechanisms into training data. Correct estimation is not a theorem about optimal data allocation; our earlier augmentation controls show why. Enlarging the hypothesis space or allowing intervention design would introduce new paths through which search and implementation affect utility. Those capabilities should be tested separately, not inferred here.

Statistical scope is finite. Independent initialization does not turn a geometry grid into a sample of all research problems. Training repeats and shared programs do not create new independent tasks. We use registered analysis units and multiplicity families, disclose incomplete service coverage, and retain unfavorable outcomes. This limits generalization while making the narrower evidence interpretable.

8 Conclusion

Research agents can improve a system while carrying an inaccurate external diagnosis. They can also benefit from accurate, transferable, revisable mechanism information. Our QEC experiments

expose both behaviors by separating accuracy from utility and intervening on the connecting record. The lesson is not to replace outcome evaluation with explanation scoring. It is to combine independent correctness checks, controlled content interventions, matched execution, and transfer or revision tests. This provides a defensible way to ask whether submitted reasons contribute to improvement, without mistaking the answer for access to private understanding.

AI USE STATEMENT

Generative AI tools were used to refine research questions and hypotheses, review literature, discuss experimental methodology, implement and test software, interpret recorded results, and draft, translate, organize, and edit this manuscript. A hosted language model is also the experimental subject (Appendix A). Synthetic circuit samples come from declared simulators, not language-model prose treated as experimental observations. Registered artifacts, programmatic audits, and retained negative outcomes ground the claims. The authors are responsible for checking the final manuscript and all AI-assisted artifacts before submission.

REPRODUCIBILITY STATEMENT

Appendices specify the physical channel, typed interface, training protocol, inferential units, and central controls. Execution records retain source/input hashes, prompts, responses, checkpoints, evaluation errors, and independent replay audits. Registrations and reports are retained locally; this manuscript does not claim an anonymous public code release has already been made. Source materials must be anonymized before submission as supplementary code.

ETHICS STATEMENT

The experiments use simulated noise and contain no human-subject or patient data. A reporting risk is overstating scientific competence from a constrained interface or selective results. We address it by separating external records from private cognition, retaining failures and null findings, and specifying scope.

REFERENCES

Johannes Bausch, Andrew W. Senior, Francisco J. H. Heras, Thomas Edlich, Alex Davies, Michael Newman, Cody Jones, Kevin Satzinger, Murphy Yuezhen Niu, Sam Blackwell, George Holland, Dvir Kafri, Juan Atalaya, Craig Gidney, Demis Hassabis, Sergio Boixo, Hartmut Neven, and Pushmeet Kohli. Learning high-accuracy error decoding for quantum processors. *Nature*, 635: 834–840, 2024. doi: 10.1038/s41586-024-08148-8. URL https://www.nature.com/articles/s41586-024-08148-8.

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

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.05080.

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

Craig Gidney. Stim: a fast stabilizer circuit simulator. *Quantum*, 5:497, 2021. doi: 10.22331/q-2021-07-06-497. URL https://arxiv.org/abs/2103.02202.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Jingjie Ning, Xiaochuan Li, Ji Zeng, Hao Kang, and Chenyan Xiong. Auto research with specialist agents develops effective and non-trivial training recipes. *arXiv preprint arXiv:2605.05724*, 2026. URL https://arxiv.org/abs/2605.05724.

A CENTRAL-STUDY PROTOCOL AND INTERPRETATION

Physical family. The central study uses a distance-3, three-round surface-code memory-Z circuit. Additional two-qubit X channels are placed on three disjoint pairs at each of three before-round locations. For a pair with coefficient α and marginal probability q,

$$P(10) = P(01) = (1-\alpha)q(1-q), \tag{2}$$

$$P(11) = q^2 + \alpha q(1-q), \tag{3}$$

$$P(00) = 1 - P(10) - P(01) - P(11). \tag{4}$$

Each qubit has marginal probability q and the pair covariance is $\alpha q(1-q)$. The registration fixes $q = .003$. Channels are independent across pair/location instances; there is no global mechanism selected once per shot. Four geometries (rows-high, rows-low, columns-high, columns-low) and six permutations of $(.1, .5, .9)$, with two independent initializations per cell, define the panel. These are held-out compositions within the declared family, not new physical laws.

Information and timing. Source counts aggregate 16,384 sampled physical-channel outcomes. The source record contains three coefficients in source-pair order, encoded in basis points. The recipient receives the source-to-target mapping for a 180-degree rotation and the coordinate convention, but not source raw counts. In observed-target regimes it receives target counts; preserved and changed prompts differ in those observations, not in a private shift label. All 48 source responses precede the 432 target responses, which precede target fitting and private evaluation. All 480 scientific responses are valid. One capacity attempt was retried under the fixed policy, giving 481 attempts; its usage is unknown, not zero.

Agent. The subject is the hosted `gpt-5.6-sol` alias, low reasoning effort, a tool-free response interface, and a 600-second deadline. At most three identical capacity-only attempts and two concurrent calls are permitted. Unfavorable scientific responses are not resampled or repaired. This configuration does not specify a service-side weight hash or establish behavior of all related model versions.

Training. Each program uses the same decoder and optimizer configuration and performs 256 updates with batch size 4,096 and learning rate 10^{-4}. Half of the synthesized examples are ordinary samples from the submitted generator; half are conditioned on at least two fired detectors. All policies share this hard-syndrome synthesis rule, including the native-distribution reference. Three seeded training repeats share initial weights and optimizer buffers within a block. Zero dropout and deterministic execution support native replay. Exact canonical program parameters define deduplication, without approximate rounding. Source-program reuse does not transfer trained source weights.

Evaluation and inference. Each block/world/repeat uses a fresh 262,144-shot panel shared across compared programs. Initial and trained error counts e_0, e_1 give $U = \log((e_0 + .5)/(e_1 + .5))$. Three repeat utilities are averaged within each of 48 blocks. Primary effects use two-sided block-level t tests, pointwise 95% intervals, and Holm correction at .05. Six-test Bonferroni simultaneous intervals are also reported. These are finite-grid parametric approximations, not distribution-free or cross-domain guarantees. Descriptive control intervals are not additional confirmatory discoveries.

Coverage and audit. The operational claim requires at least 95% valid source responses and valid responses in every target arm-by-regime cell, plus favorable corrected results for the first five primary contrasts. The sixth separately tests the homogeneous policy. All coverage fractions are one. Invalid target responses would use the homogeneous program operationally while scientific scores remain explicitly missing; no such fallback was needed here. Independent native-environment audits replay training data and checkpoint predictions. Portable local checks verify hashes and recompute summaries; they do not claim identical Stim samples across CPU architectures or software versions.

Control estimands. Mapped-source applies the geometry transform without revision. Literal source reuse preserves old geometry and coefficients and is not a one-variable understanding intervention. The count estimator uses source counts without target evidence and target counts when

supplied; it is a whole-pipeline baseline. The diagnostic oracle uses true generating coefficients, not a policy selected on the private evaluator. Homogeneous fixes all coefficients to .5, broad mixture uses an equal-weight eight-mechanism mixture, and uniform uses the native base generator under the shared synthesis rule. All controls are retained regardless of ranking.

B COMPLETE CENTRAL-STUDY RESULTS

Tables below are generated from the completed, integrity-passed analysis. They retain all primary endpoints and all prespecified regime-by-policy means and original-minus-control differences. Control intervals are descriptive and pointwise. Only the six primary contrasts carry the stated confirmatory correction.

Table 3: Central primary family: simultaneous confidence intervals and exact adjusted probabilities. All contrasts use 48 blocks.

Contrast	Mean	Simultaneous interval	Holm p
Source accuracy (MSE)	0.10445	[0.10343, 0.10547]	9.48×10^{-77}
Record value (utility)	0.00596	[0.00210, 0.00982]	1.00×10^{-4}
Record content (utility)	0.02754	[0.01705, 0.03803]	7.33×10^{-9}
Revision accuracy (MSE)	0.42404	[0.41089, 0.43719]	2.61×10^{-53}
Revision utility	0.03855	[0.02651, 0.05060]	6.31×10^{-11}
Homogeneous-policy advantage	0.00575	[0.00369, 0.00781]	2.20×10^{-9}

Table 4: Preserved mechanism, no target observations: absolute utility and original-minus-control contrasts (Δ). Intervals are descriptive pointwise 95% intervals.

Policy	U	95% interval	Δ	95% interval
Original record	0.04278	[0.03575, 0.04980]	—	—
Permuted record	0.01524	[0.00374, 0.02673]	0.02754	[0.01988, 0.03520]
No record	0.03682	[0.02973, 0.04391]	0.00596	[0.00314, 0.00878]
Mapped source (unrevised)	0.04278	[0.03575, 0.04980]	0.00000	[0.00000, 0.00000]
Literal source reuse	-0.04788	[-0.07440, -0.02137]	0.09066	[0.06668, 0.11464]
Count estimator	0.04295	[0.03582, 0.05009]	-0.00017	[-0.00211, 0.00176]
True-parameter oracle	0.04340	[0.03633, 0.05047]	-0.00062	[-0.00266, 0.00142]
Homogeneous	0.03682	[0.02973, 0.04391]	0.00596	[0.00314, 0.00878]
Broad mixture	0.02447	[0.01822, 0.03072]	0.01831	[0.01397, 0.02265]
Native reference	0.02056	[0.01575, 0.02536]	0.02222	[0.01748, 0.02697]

Table 5: Preserved mechanism, target observations: absolute utility and original-minus-control contrasts (Δ). Intervals are descriptive pointwise 95% intervals.

Policy	U	95% interval	Δ	95% interval
Original record	0.04170	[0.03436, 0.04905]	—	—
Permuted record	0.04199	[0.03459, 0.04940]	-0.00029	[-0.00184, 0.00126]
No record	0.04196	[0.03456, 0.04936]	-0.00026	[-0.00151, 0.00099]
Mapped source (unrevised)	0.04278	[0.03575, 0.04980]	-0.00107	[-0.00329, 0.00114]
Literal source reuse	-0.04788	[-0.07440, -0.02137]	0.08959	[0.06600, 0.11318]
Count estimator	0.04367	[0.03640, 0.05094]	-0.00196	[-0.00357, -0.00036]
True-parameter oracle	0.04340	[0.03633, 0.05047]	-0.00170	[-0.00350, 0.00011]
Homogeneous	0.03682	[0.02973, 0.04391]	0.00488	[0.00224, 0.00753]
Broad mixture	0.02447	[0.01822, 0.03072]	0.01723	[0.01302, 0.02144]
Native reference	0.02056	[0.01575, 0.02536]	0.02115	[0.01650, 0.02580]

Table 6: Changed mechanism, target observations: absolute utility and original-minus-control contrasts (Δ). Intervals are descriptive pointwise 95% intervals.

Policy	U	95% interval	Δ	95% interval
Original record	0.04330	[0.03453, 0.05207]	—	—
Permuted record	0.04463	[0.03530, 0.05395]	-0.00133	[-0.00321, 0.00055]
No record	0.04554	[0.03661, 0.05447]	-0.00224	[-0.00359, -0.00089]
Mapped source (unrevised)	0.00474	[-0.00825, 0.01773]	0.03855	[0.02976, 0.04735]
Literal source reuse	-0.03680	[-0.05835, -0.01526]	0.08010	[0.06356, 0.09663]
Count estimator	0.04422	[0.03512, 0.05332]	-0.00093	[-0.00271, 0.00086]
True-parameter oracle	0.04513	[0.03597, 0.05429]	-0.00184	[-0.00396, 0.00029]
Homogeneous	0.03688	[0.02829, 0.04546]	0.00642	[0.00358, 0.00926]
Broad mixture	0.02421	[0.01784, 0.03059]	0.01908	[0.01476, 0.02340]
Native reference	0.02114	[0.01587, 0.02641]	0.02215	[0.01683, 0.02747]

C SUPPORTING AND EARLIER STUDIES, INCLUDING NEGATIVE RESULTS

Local evidence and record intervention (F1). The study covers 72 tasks and four evidence views. Its primary contrasts pass their own Holm family: rich versus outcome-only joint mechanism score (mean .02077, 70 evaluable tasks, adjusted $p = 4.75 \times 10^{-26}$), eligible original versus coupling-corrupted utility (.18999, 72 tasks, $p = 1.76 \times 10^{-6}$), and rich versus outcome-only original utility (.07149, 72 tasks, $p = 1.53 \times 10^{-10}$). The serialization-placebo difference is descriptively small (.00336). The action selector also receives evidence; this is not a strict record-only channel. Three conditions have retained model failures, and missing diagnostic scores are not imputed. There are 1,829 scientific calls and 1,833 attempts, including four capacity events.

Feedback search (F2). Table 2 reports all three primary contrasts. The complete design contains 24 tasks, two evidence views, and three feedback budgets: 144 trajectories. Task-level reduction precedes inference. One model failure and one action failure remain in operational outcomes. There are 1,987 scientific calls and 1,996 attempts. At the largest budget, original-minus-corrupt utility is descriptively .13214 over 23 valid pairs; original-minus-placebo is .00317 over 46 pairs. These denominators are not pooled or substituted for the primary analysis.

Earlier memory transfer (F3). All 384 conditions completed: 32 ordered source-target pairs, two evidence views, and six memory arms. There are 1,152 scientific calls and 1,155 attempts. None of the primary memory-transfer contrasts passed its Holm family: rate-preserving shift (.09439, adjusted $p = .16491$), known placement transform (.02254, $p = .46297$), and coupling change (.15975, $p = .40203$). Positive estimates do not establish transfer. The later central study changes the task and interface and does not overwrite this result.

Changed-world forecasts (F4). Original and service-replication blocks share 48 physical tasks and archived models. Service-quota failures left the original block with only 25 valid responses in each of two coverage-critical arms; its forecast and renaming-equivalence gates remain failed. All 192 replication responses are valid. Under the frozen two-block eight-test Holm family, forecast skill is .01734 ($p = .00116$), original-minus-corrupt utility is .16343 ($p = 5.46 \times 10^{-11}$), adaptation-minus-identity-copy utility is .23049 ($p = .00204$), and renaming equivalence passes the prespecified two-one-sided-test criterion ($p = 8.43 \times 10^{-19}$; margin $\pm.0512933$). The original block supports record-content and adaptation effects but not the coverage-gated claims. Neither selective merging nor treating the blocks as 96 independent tasks is justified. Different worlds use independent evaluation draws, not individual common-noise counterfactuals. A post-analysis audit finds 180/192 original-record choices agree with MAP-hard: descriptive agreement, not internal-algorithm identification.

Original predicate-based closed loop. The initial design uses a 320-predicate physical-fault grammar and a predicate-only data compiler, rather than the later parameter interface. It completed 576 three-round episodes, 48 five-branch record-replacement blocks, and 12 finite-pool baseline blocks. Three positive primary hypotheses were not confirmed: added trace evidence for final utility (adjusted $p = .68101$), added trace evidence for support-matched rescue selectivity ($p = .53695$), and agent versus perturbed predicate utility ($p = .10809$). A broad count-two policy outperformed the trace-agent condition in its separate secondary family. This is not evidence that all information conditions are equivalent. Realized variance exceeded the planning assumption, so planned power is not reported as achieved power.

Physical rescue is not a learnable-defect oracle. A post-result audit computes the full detector/logical joint distribution of two small native circuits without sampling approximation or fault-weight truncation, using floating-point arithmetic. The native Bayes rule f_P^* has zero expected excess risk, yet many failures under a specified predicate are rescued by removing events and replaying the same rule. Rescue changes the physical input; it does not show that another syndrome-only predictor improves native expected risk. A conditional mixture $Q = (1 - \lambda)P + \lambda P(\cdot \mid H)$ can also change the Bayes decision and increase its risk under P. Accurate attribution, learnable excess risk, and useful data allocation are distinct.

Independent fixed-predicate dose and posterior-preserving controls reinforce this distinction. Reducing dose mitigates harm but does not establish an advantage over uniform training. Preserving

conditional label probabilities likewise does not guarantee finite-budget gains. These audits do not turn original agent primary tests positive. They motivate, rather than validate by definition, the central study's separate accuracy and utility measurements.

D EVIDENCE BOUNDARIES AND RESOURCES

The central analysis passes integrity checks and includes all 48 blocks, 96 execution/audit stages, and 1,983 distinct fits. Recorded successful-attempt usage is 6,100,253 input and 360,522 output tokens; one capacity-error attempt has unknown usage. Execution and audit span approximately six hours across two GPU lanes. This includes I/O, evaluation, and replay and is not GPU-active time or billing. Preparation and development are excluded from that span.

Formal manifests, call records, outputs, and native audits are retained separately. Studies are not pooled into an overall success fraction. Ongoing broader-autonomy experiments are not used in this manuscript's results or abstract; no full-cohort result is extrapolated from partial execution. The supported claim is conditional: within the supplied family and compiler, accurate external records have incremental value and evidence-driven revision improves both representation and utility. Private beliefs, natural mediation effects, necessity of understanding, and general scientific competence remain unestablished.