
ReconcileBench: Benchmarking State Reconciliation in Persistent Scientific Agents

Sai Charitha Akula
sca321@nyu.edu

Chanukya Vardhan Gujjula
crg9968@nyu.edu

Abstract

Persistent scientific agents reuse their own outputs across rounds of work. When evidence or analysis policy changes, that record can quietly become invalid: some artifacts must be updated, others must remain valid, and the effects can cross structured data, computations, prose, figures, tests, and decisions. We study the capability needed to recover validity, which we call **scientific state reconciliation**. We introduce ReconcileBench to test how well agents reconcile persistent scientific workspaces after authoritative changes. It spans controlled analytical and health repositories together with executable RNA-seq projects derived from a published computational-biology capsule, with state distributed across heterogeneous scientific artifacts including structured data, prose, tests, charts, figures, and decisions. Agents receive the existing workspace and event, but not the corrected reference or a field-level repair list. Across 106 episodes per configuration in mini-SWE-agent, the highest-scoring agent, GPT-5.6-SOL, meets its task-specific closure criterion in 62. Failures are often local-to-global: agents make the announced edit and frequently repair the immediate analysis, yet leave dependent representations or conclusions inconsistent. Reconciliation also requires semantic judgment: the same change may require one conclusion to be revised while another remains valid, and some notices call for restoring an earlier state or making no edit at all. Correctly processing new information does not guarantee that the scientific record carried into the next cycle remains valid. ReconcileBench tests whether agents can keep persistent state synchronized across representations and semantic roles before it is reused.

1 Introduction

Scientific agents increasingly work over long horizons and accumulate a persistent scientific record. They generate hypotheses, run analyses, write reports, and propose the next experiment over many steps and sessions, while harnesses preserve their tools, files, memory, and context [11, 30]. Closed-loop systems make this persistence central. Co-Scientist keeps a persistent context memory across rounds of hypothesis generation and review [10]; Robin carries the conclusions of one lab-in-the-loop round into the next [8]; and Kosmos updates a structured world model after each cycle and queries it to plan the following one [19]. The workspace becomes scientific memory: observations feed computations, computations feed figures and prose, and interpretations shape what the agent does next. This is especially consequential in computational biology, where sample metadata, analyses, figures, and experimental conclusions are carried from one research cycle into the next.

That persistent record can become stale when the underlying science changes. Evidence is corrected, samples are relabeled, analyses are revised, and decision criteria change; fixing the source does not automatically fix what was built from it. One audit found apparently mislabeled samples in 46% of the transcriptomics datasets it examined [27]. Forensic re-analysis of drug-sensitivity signatures that were already guiding treatment allocation in clinical trials found reversed sensitive/resistant

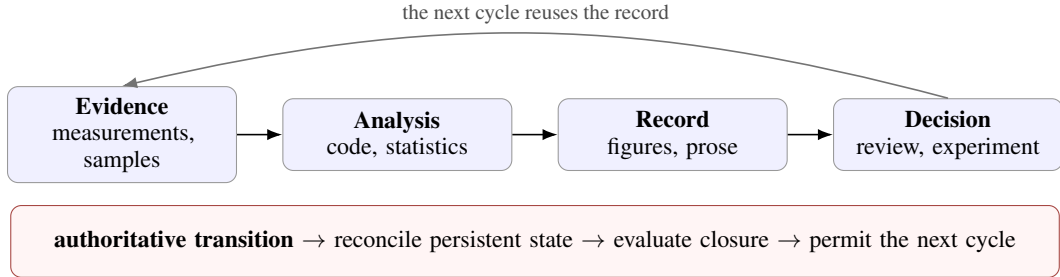


Figure 1: Scientific state reconciliation is a reliability gate before persistent state is reused.

training labels [3]. Excluding retracted trials from 3,902 meta-analyses changed the direction of the pooled effect in 8.4% of them [28], yet only 5% of the systematic reviews that had included such trials corrected their results [13]. Correcting the named value is the easy part. Agent systems inherit the problem: Robin’s preprint reported a 7.5-fold phagocytosis increase from agent analysis versus 1.75-fold from human analysis [9]; the published article reports 1.89-fold versus 1.75-fold [8]. Dependent rankings, figures, and prose built on the preprint value would need reconsideration.

Keeping the rest of the workspace valid is not simply a matter of finding references to the changed value (Figure 1). Scientific dependencies are often semantic rather than explicit: a review note restates a number from a table, a figure encodes a ranking, and a validation decision depends on a rule written in prose. The workspace may not say which numbers are observations, conventions, or fixed criteria. An agent must understand what each record means before deciding whether to update it, preserve it, restore an earlier state, or leave it alone.

We study the capability needed to recover from this drift: **scientific state reconciliation**. After an authoritative change, the agent must restore a valid persistent workspace by revising invalid records, preserving valid ones, and sometimes restoring or doing nothing. Retrieval asks whether prior state can be found; reconciliation asks whether it is still valid. Computation gives a revised quantity; reconciliation asks whether the records built on it still mean what they claim. An RNA-seq project from our benchmark shows how far apart the two can be: a correction moves one sequenced library to its true experimental condition. GPT-5.6-SOL edits the sample sheet and regenerates every contrast table exactly, yet across six such projects it repairs only 48 of 66 downstream records and fully reconciles two. In one, correcting the label shrinks the contrast’s strongest response from 14.4-fold (NR4A3) to 2.9-fold (Figure 3). Even with the revised computation in place, the dependent record may remain incomplete. Across our settings, implicitly dependent records can remain stale, fixed records may be re-derived, and restorations or no-ops can become forward edits.

ReconcileBench turns this capability into a measurable task. A generator writes an original workspace and, after one authoritative transition, its exact counterfactual from the same derivations; the agent sees only the original and the transition. Because real scientific workspaces rarely come with a known dependency graph, we follow an inspectable-to-native progression. **VisText-General** provides controlled analytical workspaces in which every dependency can be inspected, **VisText-Health** moves the same structure to health and public-health records, and **BixBench-RNA** starts from a published RNA-seq capsule, in which sample metadata and evidence feed executable contrasts, findings, and validation decisions. Our contributions are:

1. **Scientific state reconciliation as an agent capability.** We define what a persistent scientific agent must do after evidence or policy changes: revise invalid state, preserve what remains valid, and, when appropriate, restore an earlier state or make no edit.
2. **A benchmark spanning heterogeneous scientific artifacts.** ReconcileBench progresses from controlled analytical and health workspaces to executable RNA-seq projects derived from a published computational-biology capsule, following authoritative changes across structured data, analyses, prose, tests, figures and charts, and scientific decisions. Original and counterfactual workspaces come from the same derivations, allowing us to check what must change and what must remain valid.
3. **Evidence of a local-to-global gap.** Agents can correct a source and even reproduce an immediate scientific computation exactly while leaving downstream records stale. The

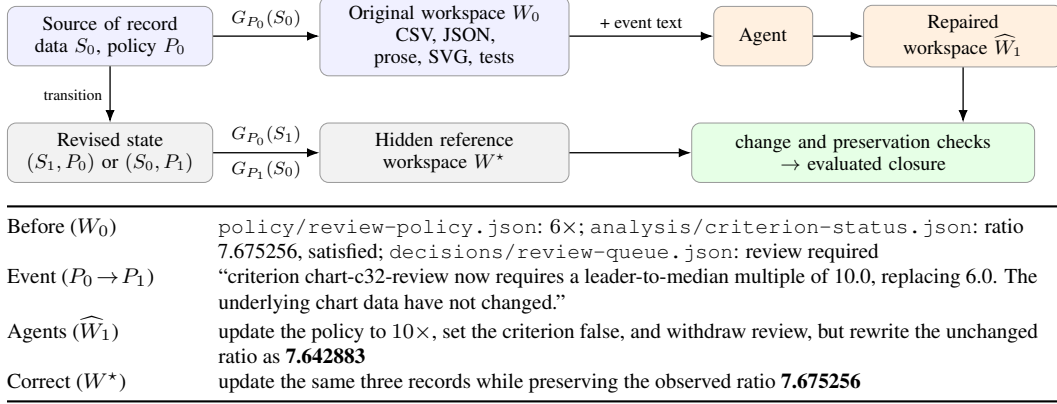


Figure 2: **Top:** the source-correction reference is $G_{P_0}(S_1)$ and the policy-revision reference is $G_{P_1}(S_0)$. Agents receive file-family guidance, but not the corrected reference or field-level grading targets. **Bottom:** in one policy transition, the three GPT-5.6 agents repair the criterion truth and review decision but re-derive an observed ratio that the event explicitly holds fixed. The decisions are locally correct, yet none of the three workspaces is fully reconciled.

observed failures identify requirements for propagation, preservation, restoration, and abstention before the workspace is reused.

We will release the benchmark codebase, workspace generators, graders, and evaluation artifacts.

2 Scientific state reconciliation

2.1 State, policy, and authoritative transitions

Let S be authoritative scientific state and P workspace policy: derivation conventions, analysis criteria, authority relations, and revision rules. An external deterministic generator G_P yields an intact workspace $W = G_P(S)$. A source correction changes $(S_0, P_0) \rightarrow (S_1, P_0)$ and has target $W^* = G_{P_0}(S_1)$; a criterion revision changes $(S_0, P_0) \rightarrow (S_0, P_1)$ and has target $W^* = G_{P_1}(S_0)$. A withdrawn correction returns to the earlier valid state; an already-applied notice leaves the target workspace unchanged.

An authoritative *transition* changes the underlying evidence or policy; the *event* is the notice the agent receives about that change. The agent works in a copy of W_0 and receives the event during the episode. VisText’s README explicitly lists source-to-file-family dependencies; data/charts.csv and analysis/INDEX.md connect chart IDs to workstream codes through chart titles. The generator G_P , corrected workspace W^* , and field-level grading manifest are withheld. The agent must use the available guidance to identify the specific records affected and determine whether each needs revision or preservation.

An episode reaches *closure* when every check in its task-specific validity boundary passes in the same final workspace. Records invalidated by the transition must be revised, while records it leaves valid must remain so. Across tasks, the evaluation covers source, computation, representation, and conclusion records; it does not inspect every file. Many correct local edits cannot compensate for one stale or wrongly altered evaluated record.

3 Benchmark design

3.1 Scientific workspaces and transitions

ReconcileBench uses three scientific workspace constructions. **VisText-General** provides controlled analytical repositories with inspectable dependencies across tables, prose, tests, and figures. **VisText-Health** uses the same construction for health and public-health records. **BixBench-RNA** begins with a published RNA-seq capsule and connects experimental metadata to executable con-

Table 1: ReconcileBench workspaces and authoritative transitions. Δ and K list state affected by each transition; §3.3 defines the task-specific closure checks.

Setting and transition	Episodes	Authoritative change	Δ : state to update	K: state to keep valid
VisText-General: evidence correction	20	One chart source value overwritten or deleted	Source, affected calculations, ranks, dashboard, test, SVG, criterion truth, review flag	Fixed criterion, dated history, unaffected calculations
VisText-General: review-rule revision	20	Review criterion revised above the unchanged observed ratio	Rule, criterion truth, review flag	Source data and observed ratio
VisText-Health: evidence correction	14	Health/public-health chart source value overwritten or deleted	Source, affected calculations, ranks, summary, figure	Dated history and unaffected summaries
BixBench-RNA: condition-label correction	6	One sequenced library’s experimental condition corrected; raw counts fixed	Sample sheet, contrast, ranks, figure, dashboard, test, summary	Raw counts and scored control/history
BixBench-RNA: mixed evidence/policy	6	Library relabel, evidence withdrawal, support-rule revision, or compound transition	Affected support counts/joins and findings/actions licensed by revised support	Surviving evidence, fixed policy/data, history, and still-supported findings
VisText-General: prose evidence correction	20	One chart source value corrected; criterion number unchanged in the reference	Lapsed criterion claims (18) and dependent review flags (17)	Reference keeps criterion fixed, but ordinary workspace leaves its status implicit
VisText-General: prose criterion revision	20	Review criterion changes; source data and observed ratio stay fixed	Criterion-truth sentence and dependent review-flag sentence (40 scored statements)	Source data and observed ratio, fixed by the event but not separately scored

Table 2: Semantic layers of scientific state and their required operations. Direct state, computation (C1), relation or representation (C2), and conclusion or action (C3) may each require revision (Δ) or preservation (K).

Layer	Δ : affected state must change	K: state must remain valid
Direct source/rule	Correct chart row or RNA label; withdraw evidence; revise threshold	Keep source data fixed during a rule change, or the rule fixed during a data correction
C1 computation	Recompute mean, ratio, contrast quantity, or support count	Keep observed ratio during a rule change or an unaffected count
C2 relation/representation	Repair leader, rank, evidence join, assertion, dashboard, or SVG	Keep an unaffected relation or figure valid
C3 conclusion/action	Revise criterion truth, review flag, finding, or action when the transition requires it	Keep a dated or still-supported conclusion/action valid

trasts and scientific decisions. Table 1 states what each authoritative transition requires the agent to update and what it must keep valid.

3.2 Semantic class and required operation

We check the announced source or rule revision directly, then follow its dependent records. Each scored downstream field is classified by semantic role: C1 computation, C2 relation or representation, or C3 conclusion or action. An independent operation label specifies whether the transition requires a change (Δ) or preservation (K, for *keep valid*). Thus a changed ratio is C1- Δ ; when only the review rule changes, that same observed ratio is C1-K. A finding is C3- Δ when support no longer licenses it, but C3-K when independent support still suffices.

Mixed records receive the deepest class needed to grade their target. Dashboards, executable assertions, and concentration records combine arithmetic with selection or joins, so they are C2. An SVG can combine labels, ordering, values, and geometry; the vector check tests painted bar count, order, relative lengths, and one numeric leader label. Preservation follows each record’s semantic or exact check rather than requiring every valid file to be byte-identical. The Δ /K labels describe record-level requirements; restoration and no-op also test which operation the episode itself calls for (§4.3).

3.3 Construction, units, and grading

Original and corrected workspaces. The same derivation code builds an original workspace and a hidden corrected reference from the authoritative state before and after each transition. The agent sees the original files and the event, but not that code or the reference. VisText-General draws correction values from nine published series [25]; each of its 20 project workspaces catalogs 30 data series and contains a median of 298 files. VisText-Health uses seven health/public-health series; 14 of its 16 generated references are scored after excluding two with stale harvested captions. BixBench-RNA uses a published experimental capsule [18] in two complementary sets of cases (Table 1).

Episodes and required checks. A *workspace* is one persistent project file tree; an *episode* is one model run on it under an authoritative transition. VisText-General reuses 20 templates for evidence correction, review-rule revision, and two conclusion-maintenance conditions. VisText-Health contributes 14 evidence corrections; BixBench-RNA contributes six condition-label and six mixed evidence/policy transitions. Thus each model has 66 propagation and 40 conclusion-maintenance episodes, or 106 runs in total. Each field check specifies an operation: Δ requires the corrected value or claim, while K requires a still-valid record to remain valid. Specified unauthorized collateral edits are checked separately.

Agent and harness setup. Each episode runs in a copy of the original workspace; the agent receives the event at the relevant turn. With mini-SWE-agent’s shell interface [30], it can read and write files and run scripts or tests. The VisText README and index give substantial dependency guidance, but the agent has no corrected reference, field-level target list, or structured-patch tool. The sandbox permits edits to this workspace but excludes gold, the grader, and other episodes; Docker disables networking, while local runs use filesystem isolation and scrubbed credentials. We evaluate six configurations from two model families in mini-SWE-agent: GPT-5.6-LUNA, TERRA, and SOL at low reasoning [22], and Claude Haiku 4.5 without extended thinking plus Sonnet 5 and Opus 5 at low effort. We also evaluate GPT-5.6-SOL in Codex CLI and Claude Opus 5 in Claude Code on the same tasks (Appendix A). Mini-SWE-agent allows at most 12 agent steps per turn and 60 per episode, increased to 400 for retraction episodes. Responses are capped at 4,096 tokens, and episodes have a one-hour wall limit.

Scoring. We use two complementary graders:

- **Deterministic checks** compare source rows, JSON values, ordered lists, selected painted SVG bar geometry and its numeric leader label, and executable tests with the reference.
- **Prose proposition checks** use GPT-5.6-SOL as a low-reasoning LLM judge against code-derived targets. For propagation prose it sees the transition and full before/after files; for C3 conclusion prose it sees selected criterion and review-flag sentences with the relevant values and rule. Neither receives reference wording (Appendix B).

Closure depends on the transition being tested. In the 66 propagation episodes, the announced source or rule, every affected C1 computation and C2 representation, and their preservation checks must all pass in the same final workspace. In the 40 conclusion-maintenance episodes, both the criterion claim and its dependent review flag must be correct, whether that requires revision or preservation. The overall result is the number of closed episodes across these two groups, out of 106. Retraction and already-applied episodes are reported separately because they test different episode-level operations.

4 Evaluation across scientific workspaces

In mini-SWE-agent, GPT-5.6-SOL closes the most workspaces: 62/106 (58.5%). Correcting the announced source or rule is usually not the obstacle; failures emerge in the records that depend on it or must remain unchanged.

Table 3: Propagation results across scientific workspaces. Field scores show whether agents update or preserve Direct, C1, and C2 state; closure requires their applicable checks in an episode to pass together.

Transition	Agent	Direct		C1 compute		C2 relation/figure		Closure
		Δ	K	Δ	K	Δ	K	
VisText-General								
Evidence	GPT-5.6-LUNA	20/20	20/20	38/40	–	137/176	3/4	0/20
	GPT-5.6-TERRA	20/20	20/20	37/40	–	162/176	4/4	6/20
	GPT-5.6-SOL	20/20	20/20	39/40	–	166/176	4/4	9/20
Rule	GPT-5.6-LUNA	20/20	–	–	14/20	–	–	14/20
	GPT-5.6-TERRA	20/20	–	–	12/20	–	–	12/20
	GPT-5.6-SOL	20/20	–	–	17/20	–	–	17/20
VisText-Health								
Evidence	GPT-5.6-LUNA	13/14	–	19/26	0/2	71/122	3/4	0/14
	GPT-5.6-TERRA	13/14	–	24/26	2/2	102/122	4/4	2/14
	GPT-5.6-SOL	14/14	–	26/26	2/2	113/122	4/4	6/14
BixBench-RNA								
Metadata	GPT-5.6-LUNA	6/6	–	4/12	–	15/54	–	0/6
	GPT-5.6-TERRA	6/6	–	10/12	–	41/54	–	1/6
	GPT-5.6-SOL	6/6	–	10/12	–	38/54	–	2/6
Mixed	GPT-5.6-LUNA	8/8	6/6	6/10	10/10	9/12	6/6	4/6
	GPT-5.6-TERRA	8/8	6/6	10/10	10/10	8/12	6/6	2/6
	GPT-5.6-SOL	8/8	6/6	10/10	10/10	11/12	6/6	5/6

Table 4: Prose conclusion-maintenance results on VisText-General across GPT-5.6 and Claude variants. Episode closure requires the criterion claim and dependent review flag to be correct together.

Transition	GPT-5.6 family			Claude family		
	LUNA	TERRA	SOL	Haiku 4.5	Sonnet 5	Opus 5
VisText-General: prose evidence correction	4/35	12/35	16/35	0/35	5/35	14/35
Criterion claim	0/18	0/18	2/18	0/18	0/18	2/18
Dependent review flag	4/17	12/17	14/17	0/17	5/17	12/17
Episode closure	1/20	1/20	3/20	2/20	2/20	4/20
VisText-General: prose criterion revision	6/40	35/40	40/40	0/40	27/40	37/40
Episode closure	3/20	16/20	20/20	0/20	10/20	18/20
C3 prose closure, both transitions	4/40	17/40	23/40	2/40	12/40	22/40

Table 5: Overall reconciliation performance across GPT-5.6 and Claude models. Field scores locate errors; episode closure requires the task-specific checks to pass in the same final workspace.

Agent	Field checks (correct/applicable)						Episode closure		
	Direct Δ	C1 Δ	C1 K	C2 Δ	C2 K	C3 prose Δ	Propagation	Conclusion	Overall
GPT-5.6-LUNA	67/68	67/88	24/32	232/364	12/14	10/75	18/66	4/40	22/106 (20.8%)
GPT-5.6-TERRA	67/68	81/88	24/32	313/364	14/14	47/75	23/66	17/40	40/106 (37.7%)
GPT-5.6-SOL	68/68	85/88	29/32	328/364	14/14	56/75	39/66	23/40	62/106 (58.5%)
Claude Haiku 4.5 (no thinking)	49/68	16/88	23/32	52/364	14/14	0/75	8/66	2/40	10/106 (9.4%)
Claude Sonnet 5	60/68	33/88	20/32	66/364	14/14	32/75	11/66	12/40	23/106 (21.7%)
Claude Opus 5	68/68	60/88	22/32	185/364	14/14	51/75	25/66	22/40	47/106 (44.3%)

4.1 Correct sources, stale representations

Correcting the announced source is often reliable; propagating that correction through C1 computations and C2 representations is less so. In VisText-General, GPT-5.6-SOL applies all 20 source corrections but closes only 9/20 propagation episodes. Most remaining discrepancies are in C2 representations, especially figures.

The same pattern appears in health and public-health analytical records. GPT-5.6-SOL repairs 139/148 downstream records but closes only 6/14 VisText-Health workspaces. High local accuracy can still leave the scientific record inconsistent. Under this turn budget, Claude Opus 5 corrects

every source in VisText-General and VisText-Health, yet closes only 3/20 and 5/14 workspaces, respectively.

Exact computation, stale scientific record. Six project workspaces generated from a published BixBench RNA-seq capsule each contain a sample sheet, raw counts, executable contrast, rankings, dashboard, figure, tests, and scientific notes [18]. The capsule measures 63,677 genes in 30 sequenced libraries. One library, identified as 6-3, is reassigned from the `Cisplatin_IC50` condition to `Cisplatin_IC50_Serum_starvation_16h`; its gene counts do not change. The new label changes which libraries enter the contrast, so the result and dependent records must be reconsidered. Appendix C shows the exact row, contrast, project files, and scored descendants.

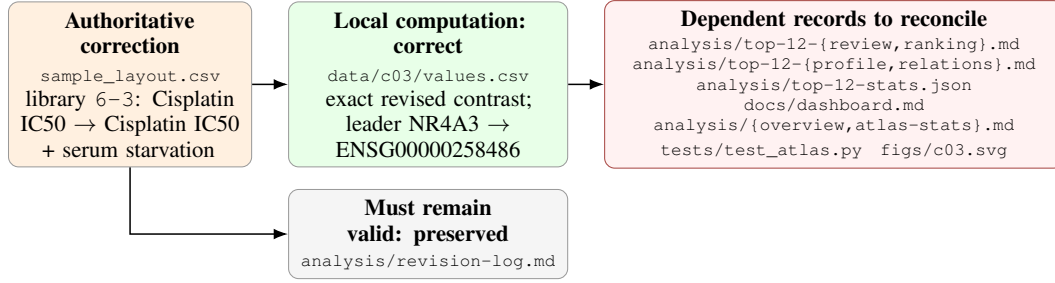


Figure 3: An RNA-seq reconciliation episode. GPT-5.6-SOL corrects the sample label and produces the byte-exact revised contrast, but the dependent scientific record does not reach closure; unrelated dated history stays valid.

All three GPT-5.6 configurations correct every sample label and regenerate every revised contrast byte-for-byte, yet downstream state remains stale. Even GPT-5.6-SOL closes only 2/6 projects. Figure 3 shows why the exact local computation is only one step: its result must also be reflected in the analyses, dashboard, tests, figure, and scientific notes that persist in the workspace. Claude Opus 5 also closes only 4/6 RNA projects.

4.2 Conclusions must change selectively

A change in evidence or policy can alter what the workspace is justified in concluding, but not every supporting record should change with it. In the General rule-revision task, the review threshold becomes stricter while the chart data and recorded leader-to-median ratio stay fixed. The repair is selective: the agent should update whether the criterion is met and the resulting review decision, while leaving the measurement untouched. The three GPT-5.6 agents reach the right decision, but some replace the fixed ratio with a newly computed value under a different median convention (Figure 2). The conclusion is correct, yet its supporting record has drifted; even GPT-5.6-SOL closes only 17/20 workspaces. Claude Opus 5 also closes only 10/20 rule-revision workspaces.

The prose tasks test the same idea in written conclusions (Table 4). When the rule changes but the evidence remains fixed, the agents must revise the criterion and review statements accordingly; performance ranges from 6/40 correct statements for GPT-5.6-LUNA to 40/40 for GPT-5.6-SOL. The complementary condition changes the evidence while keeping the numerical criterion fixed in the reference; GPT-5.6-SOL repairs 16/35 statements. This condition also asks the agent to infer which numerical criterion still governs the claim.

The RNA-seq withdrawal cases make selective conclusion maintenance concrete. Two independent experiments, E1 and E2, can support the same candidate gene. When E1 is withdrawn after failing quality control, AHRR still has support from E2, so its `SUPPORTED` finding and `RUN_VALIDATION` action should remain unchanged. CYP1A1 has no qualifying support left, so its finding should change to `NOT_ESTABLISHED` and its action to `HOLD`. The three GPT-5.6 agents make this distinction correctly. The same withdrawal can therefore preserve one conclusion while revising another, as shown in Appendix Figure 4.

4.3 Choosing the reconciliation operation

Not every change calls for another forward update. Sometimes the correct response is to return to an earlier valid state; sometimes the workspace is already correct and should be left alone.

We test both situations. In the retraction setting, an earlier correction is withdrawn after 2, 6, or 12 intervening turns. The agent must recover the workspace that was valid before the correction, including its dependent analyses, figures, tests, and summaries. The three GPT-5.6 configurations restore many individual files but no complete workspace: some files retain the temporary corrected state or become *third states* that match neither version.

The already-applied setting tests the opposite behavior. The announced correction is already present, so the right response is no edit. Yet all three GPT-5.6 agents sometimes alter valid state; even GPT-5.6-SOL leaves only 10/20 workspaces untouched. Claude Opus 5 also closes only 2/60 retractions and leaves 13/20 already-correct workspaces untouched. Before editing, an agent must decide whether the transition calls for updating, preserving, restoring, or doing nothing.

Table 6: Restoration and abstention results in mini-SWE-agent. Success requires recovering the evaluated earlier workspace or leaving an already-correct workspace unchanged; file-level outcomes appear in Appendix Table 9.

Agent	Restoration	Abstention
GPT-5.6-LUNA	0/60	1/20
GPT-5.6-TERRA	0/60	3/20
GPT-5.6-SOL	0/60	10/20
Claude Haiku 4.5	1/60	0/20
Claude Sonnet 5	11/60	8/20
Claude Opus 5	2/60	13/20

Using Codex CLI and Claude Code improves closure on the main tasks; Opus 5 rises from 47/106 to 67/106, although conclusion-maintenance, restoration, and abstention failures remain (Appendix Table 7).

5 Implications for scientific-agent design

Persistent scientific agents reuse their own outputs. A stale figure, summary, or decision can therefore become input to the next round of reasoning: an agent may recompute an analysis correctly, then read an outdated explanation or plan from an obsolete finding. We do not measure such downstream cycles, but the failures here identify a state-validity risk. Before previous work is reused, an agent needs to know which parts of its record still support the claims and actions built on them.

Avoiding that risk requires more than knowing that two artifacts are connected. The agent must understand the *semantic role* of each record and how its meaning is expressed across representations. A value may be an observation, a derived quantity, a fixed criterion, or evidence for a decision; the same dependency can therefore require revision under one transition and preservation under another (§4.2). Tables, prose, tests, dashboards, and figures are different representations of scientific state, not independent files. Reconciliation requires semantic lineage across modalities: what each artifact represents and what continues to justify it [7].

Mechanically derived artifacts can be regenerated from structured state [20], while prose, figures, and decisions can retain provenance linking them to the evidence, computation, or policy that justifies them. A harness can also maintain versioned state, track which authoritative transitions have been applied, and expose instance-level lineage. Before persistent state is reused, a reconciliation check can report what changed, what was preserved, and what remains unresolved. These are design implications, not interventions tested here.

6 Related work

Scientific agents and fixed-start benchmarks. Scientific agents analyze biomedical data and carry findings across research cycles [2, 8, 10, 11]. Biology benchmarks chiefly grade forward work from fixed inputs: analysis questions [18], analysis programs [5], reproducibility [21, 24], and study-scale or long-horizon pipelines [6, 15]. ReconcileBench instead begins with a written scientific record and asks whether it remains valid when its evidence or policy changes.

Scientific visualization and multimodal records. ChartAgent reasons over chart images [14]; SciVisAgentBench evaluates scientific analysis and visualization agents [1]; and Beaver curates evidence distributed across text, tables, and figures in papers [31]. Our question is different: does an *existing* visual record remain synchronized after the underlying scientific state changes? Figures and charts are checked alongside data, prose, tests, and decisions. We evaluate maintenance of visual artifacts, not image-based perception by the agent.

Workspace dependencies and repair. Agent evaluations show that the scaffold can affect both accuracy and cost [12]. Workspace-Bench evaluates agents in large heterogeneous file workspaces with explicit and implicit dependencies [26]. MemoRepair repairs derived memory using influence provenance [32]; REVISE uses recorded data and control dependencies in a workflow to retain or recompute valid work [23]. ReconcileBench gives file-family guidance but withholds the corrected reference and field-level repair list. Agents must identify both affected records and their semantic roles: the same connection can require recomputation after an evidence change but preservation after a rule change.

Change propagation and state evaluation. EditPropBench propagates factual edits through dependent claims in synthetic scientific manuscripts, including implicit textual consequences [16]. Multi-turn report-revision agents also regress on previously covered content while addressing new feedback [4]. ReconcileBench extends this problem to executable, structured, visual, and decision records and to policy as well as evidence transitions. LongDS-Bench tests evolving analytical state and rollback but grades later queries rather than repair of existing artifacts [29]. AgentAbstain tests broad reasons not to act [17]; our no-op requires recognizing that a legitimate update is already present in the persistent workspace.

7 Limitations and future work

ReconcileBench uses controlled, relatively small workspaces: generated VisText repositories and executable analyses from one public RNA-seq capsule. Most episodes introduce one transition; one RNA case combines several. Longer projects, overlapping changes, and unannounced updates remain outside this evaluation. We evaluate three configurations per model family in mini-SWE-agent, plus SOL in Codex CLI and Opus 5 in Claude Code, without estimating run-to-run variability. Prose uses one model-based proposition judge; alternative judges and broader human validation would strengthen these measurements.

Future evaluations should vary model families, reasoning settings, frameworks, and interfaces, treating the *model and harness together* as the scientific agent. Harder tasks could add external file edits, notebook reruns, regenerated figures, interacting corrections, conflicts, branches, and delayed withdrawals.

Finally, continuing agents from stale and reconciled workspaces would measure how missed maintenance changes later analyses or experiment choices. Independently authored projects, more domains and real errata, repeated runs, and scientist review of targets, prose, figures, and reuse safety would make the benchmark broader and more rigorous.

8 Conclusion

Persistent scientific agents can make the announced edit and reproduce the immediate computation while leaving downstream state stale, changing valid records, or editing when restoration or no action is required. Closing this gap requires semantic lineage and the correct operation, followed by a closure check before reuse.

References

- [1] Kuangshi Ai, Haichao Miao, Kaiyuan Tang, Nathaniel Gorski, Jianxin Sun, Guoxi Liu, Helgi I. Ingolfsson, David Lenz, Hanqi Guo, Hongfeng Yu, Teja Leburu, Michael Molash, Bei Wang, Tom Peterka, Chaoli Wang, and Shusen Liu. SciVisAgentBench: A benchmark for evaluating scientific data analysis and visualization agents. *arXiv preprint arXiv:2603.29139*, 2026. URL <https://arxiv.org/abs/2603.29139>.
- [2] Samuel Alber, Bowen Chen, Eric Sun, Alina Isakova, Aaron J. Wilk, and James Zou. CellVoyager: AI CompBio agent generates new insights by autonomously analyzing biological data. *Nature Methods*, 23(4):749–759, 2026. doi: 10.1038/s41592-026-03029-6.
- [3] Keith A. Baggerly and Kevin R. Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, 3(4):1309–1334, 2009. doi: 10.1214/09-AOAS291.
- [4] Bingsen Chen, Boyan Li, Ping Nie, Yuyu Zhang, Xi Ye, and Chen Zhao. Beyond single-shot writing: Deep research agents are unreliable at multi-turn report revision. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13325–13356, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.609.
- [5] 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 *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL <https://openreview.net/forum?id=6z4YKr0GK6>.
- [6] Ian Diks, Zhen Yang, Arjun Banerjee, Tim Proctor, and Kenny Workman. scBench-Long: Verifiable benchmarking of long-horizon single-cell biology. *arXiv preprint arXiv:2606.26563*, 2026.
- [7] Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272, 1979. doi: 10.1016/0004-3702(79)90008-0.
- [8] Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodriques. A multi-agent system for automating scientific discovery. *Nature*, 655(8122):497–505, 2026. doi: 10.1038/s41586-026-10652-y.
- [9] Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Jon M. Laurent, Muhammed T. Razzak, Andrew D. White, Michaela M. Hinks, and Samuel G. Rodriques. Robin: A multi-agent system for automating scientific discovery. *arXiv preprint arXiv:2505.13400*, 2025. URL <https://arxiv.org/abs/2505.13400>.
- [10] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillion, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, 2026. doi: 10.1038/s41586-026-10644-y.
- [11] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Ryan Li, Yusuf Roohani, Lin Qiu, Shiyi Cao, Gavin Li, Junze Zhang, Di Yin, Rick Wierenga, Deniz Kavi, Sherry Liu, Tianwei She, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew T. Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael P. Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Autonomous biomedical research with an artificial intelligence agent. *Science*, 393(6813):eadz4351, 2026. doi: 10.1126/science.adz4351.
- [12] Sayash Kapoor, Benedikt Stroebel, Peter Kirgis, Nitya Nadgir, Zachary S Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndjomga, Dheeraj Oruganty, Sophie Luskin, Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, Juyong Lee, Tengjun Jin, Yifan Mai, Yifei Zhou, Yuxuan Zhu, Rishi Bommasani, Daniel Kang, Dawn Song, Peter Henderson, Yu Su, Percy Liang, and Arvind Narayanan. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. In *International Conference on Learning Representations*, 2026.
- [13] Yuki Kataoka, Masahiro Banno, Yasushi Tsujimoto, Takashi Ariie, Shunsuke Taito, Tomoharu Suzuki, Shiho Oide, and Toshi A. Furukawa. Retracted randomized controlled trials were cited and not corrected in systematic reviews and clinical practice guidelines. *Journal of Clinical Epidemiology*, 150:90–97, 2022. doi: 10.1016/j.jclinepi.2022.06.015.
- [14] Rachneet Kaur, Nishan Srishankar, Zhen Zeng, and Sumitra Ganesh. ChartAgent: A multimodal agent for visually grounded reasoning in complex chart question answering. In *Proceedings of the 64th An-*

- nual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18488–18555, 2026. doi: 10.18653/v1/2026.acl-long.843. URL <https://aclanthology.org/2026.acl-long.843/>.
- [15] Zane Koch, Asmamaw T. Wassie, Javier Valdes-Aleman, Jason Lee, Michaela M. Hinks, Samuel G. Rodrigues, Andrew D. White, and Jon M. Laurent. BixBench3: Benchmarking AI agents on research-study-scale computational biology tasks. *arXiv preprint arXiv:2608.25286*, 2026. URL <https://arxiv.org/abs/2608.25286>.
 - [16] Garvin Kruthof. EditPropBench: Measuring factual edit propagation in scientific manuscripts. *arXiv preprint arXiv:2605.02083*, 2026.
 - [17] Xun Liu, Yi Evie Zhang, Vira Kasprova, Parisa Rabbani, Pardis Sadat Zahraei, Tianyu Zhang, Ali Ebrahimpour-Borojeny, and Varun Chandrasekaran. AgentAbstain: Do LLM agents know when not to act? *arXiv preprint arXiv:2607.10059*, 2026.
 - [18] Ludovico Mitchener, Jon M. Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan, Geemi P. Wellawatte, Andrew White, Lorenzo Sani, and Samuel G. Rodrigues. BixBench: A comprehensive benchmark for LLM-based agents in computational biology. *arXiv preprint arXiv:2503.00096*, 2025.
 - [19] Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C. Landsness, Daniel L. Barabasi, Siddharth Narayanan, Nicky Evans, Shriya Reddy, Martha Foiani, Aizad Kamal, Leah P. Shriver, Fang Cao, Asmamaw T. Wassie, Jon M. Laurent, Edwin Melville-Green, Mayk Caldas, Albert Bou, Kaleigh F. Roberts, Sladjana Zagorac, Timothy C. Orr, Miranda E. Orr, Kevin J. Zvezdaryk, Ali E. Ghareeb, Laurie McCoy, Bruna Gomes, Euan A. Ashley, Karen E. Duff, Tonio Buonassisi, Tom Rainforth, Randall J. Bateman, Michael Skarlinski, Samuel G. Rodrigues, Michaela M. Hinks, and Andrew D. White. Kosmos: An AI scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025. URL <https://arxiv.org/abs/2511.02824>.
 - [20] Andrey Mokhov, Neil Mitchell, and Simon Peyton Jones. Build systems à la carte. *Proceedings of the ACM on Programming Languages*, 2(ICFP):1–29, 2018. doi: 10.1145/3236774.
 - [21] Nitya Nadgir, Sayash Kapoor, Kangheng Liu, Peter Kirgis, Matilda Orona, Stephan Rabanser, Tilman Bayer, Abhishek Shetty, Yue Ling, Derrick Chan-Sew, Rumi Nakagawa, Saiteja Utpala, Zachary S. Siegel, and Arvind Narayanan. Life after benchmark saturation: A case study of CORE-Bench. *arXiv preprint arXiv:2606.26158*, 2026.
 - [22] OpenAI. Using GPT-5.6. OpenAI API model guidance, 2026. URL <https://developers.openai.com/api/docs/guides/latest-model?model=gpt-5.6>.
 - [23] Ruoling Qi, Xuaner Wu, Penghang Liu, Jian Chen, and Yirui Liu. REVISE: Validity-guided recovery for online revisions in agent workflows. *arXiv preprint arXiv:2609.00643*, 2026.
 - [24] Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL <https://openreview.net/forum?id=BsMMc4MEGS>.
 - [25] Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. VisText: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298, 2023. doi: 10.18653/v1/2023.acl-long.401.
 - [26] Zirui Tang, Xuanhe Zhou, Yumou Liu, Linchun Li, Yukai Wu, Weizheng Wang, Hongzhang Huang, Wei Zhou, Jun Zhou, Jiachen Song, Shaoli Yu, Jinqi Wang, Zihang Zhou, Hongyi Zhou, Yuting Lv, Jinyang Li, Jiashuo Liu, Ruoyu Chen, Chunwei Liu, Guoliang Li, Jihua Kang, and Fan Wu. Workspace-bench 1.0: Benchmarking AI agents on workspace tasks with large-scale file dependencies. *arXiv preprint arXiv:2605.03596*, 2026.
 - [27] Lilah Toker, Min Feng, and Paul Pavlidis. Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies. *F1000Research*, 5:2103, 2016. doi: 10.12688/f1000research.9471.2. Version 2.
 - [28] Chang Xu, Shiqi Fan, Yuan Tian, Fuchen Liu, Luis Furuya-Kanamori, Justin Clark, Chao Zhang, Sheng Li, Lifeng Lin, Haitao Chu, Sheyu Li, Su Golder, Yoon Loke, Sunita Vohra, Paul Glasziou, Suhail A Doi, and Hui Liu. Investigating the impact of trial retractions on the healthcare evidence ecosystem (VITALITY study i): retrospective cohort study. *BMJ*, 389:e082068, 2025. doi: 10.1136/bmj-2024-082068.
 - [29] Kewei Xu, Xiaoben Lu, Shuofei Qiao, Zihan Ding, Haoming Xu, Lei Liang, and Ningyu Zhang. LongDS-Bench: On the failure of long-horizon agentic data analysis. *arXiv preprint arXiv:2605.30434*, 2026. URL <https://arxiv.org/abs/2605.30434>.
 - [30] John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1601.
 - [31] Sheng Zhang, Qin Liu, Renqian Luo, Shufang Xie, Reuben Tan, Sean Hayes, Gregory Bryman, Wendong Ge, Ruilian Zhang, Oluwaseun Egbelowo, Kelly Yee, and Hoifung Poon. Building agent harnesses for scientific curation from multimodal sources. *arXiv preprint arXiv:2606.21005*, 2026. URL <https://arxiv.org/abs/2606.21005>.

- [32] Yang Zhao, Chengxiao Dai, Mengying Kou, and Yue Xiu. MEMOREPAIR: Barrier-first cascade repair in agentic memory. *arXiv preprint arXiv:2605.07242*, 2026.

A Matched harness comparison

We compare GPT-5.6-SOL in mini-SWE-agent and Codex CLI, and Claude Opus 5 in mini-SWE-agent and Claude Code. Each model-harness pair faces the same 106 main episodes and scoring rules. Codex CLI and Claude Code change the prompts and tools and do not impose mini-SWE-agent’s 12-step per-turn cap. The last two rows report separate already-applied and retraction tasks, with 20 and 60 episodes respectively.

Across the 106 main episodes, the Opus 5 gain favors Claude Code in 26 discordant pairs versus six for mini-SWE-agent (two-sided paired sign test, $p = 0.0005$).

Table 7: Episode success for matched models in mini-SWE-agent, Codex CLI, and Claude Code. The final two rows report separate already-applied and retraction tasks, with 20 and 60 episodes respectively; neither enters the 106 main episodes.

Setting	GPT-5.6-SOL		Claude Opus 5	
	mini-SWE	Codex CLI	mini-SWE	Claude Code
VisText-General evidence	9/20	9/20	3/20	13/20
VisText-General rule	17/20	17/20	10/20	10/20
VisText-Health evidence	6/14	11/14	5/14	13/14
BixBench-RNA metadata	2/6	5/6	4/6	5/6
BixBench-RNA mixed	5/6	4/6	3/6	4/6
VisText-General prose evidence	3/20	2/20	4/20	2/20
VisText-General prose criterion	20/20	18/20	18/20	20/20
Main-episode closure	62/106	66/106	47/106	67/106
Already applied (no edit)	10/20	4/20	13/20	8/20
Retraction (restore)	0/60	0/60	2/60	2/60

B Prose judging

Structured source values, quantities, ordered relations, and exact decisions use deterministic checks. Free-form prose is evaluated against code-derived propositions by GPT-5.6-SOL at low reasoning effort, one vote per site, without a temperature or top- p override. For propagation prose, the judge receives the event, corrected rows, derivation, target proposition, presence policy, and full before/after files. For C3 conclusion prose, it receives selected criterion and review-flag sentences with the relevant evidence and rule values. Neither receives reference wording.

C BixBench-RNA workspace walkthrough

What the source data look like. The capsule contains a raw count matrix with one row per gene and one column per sequenced library, a gene-identifier map, and a sample sheet. The sample sheet supplies the experimental condition used to group libraries for contrast computation. One corrected row is:

Source file	Library ID	Original condition	Corrected condition
data/sample_layout.csv	6-3	Cisplatin_IC50	Cisplatin_IC50_Serum_starvation_16h

The corresponding column of data/counts_raw_unfiltered.csv is unchanged. Re-executing the fixed counts-per-million analysis changes the Cisplatin_IC50_vs_Untreated contrast leader from NR4A3 at 14.446-fold to ENSG00000258486 at 2.852-fold.

Downstream records. Beyond the contrast table, we evaluate how the revised result is represented in these persistent records:

Table 8: Downstream records evaluated after an RNA condition-label correction. The revised contrast must be reflected in analyses, dashboard, tests, figure, and scientific summaries.

Workspace record	Representation	Required synchronization
analysis/*-review.md	prose	leading gene and value
analysis/*-ranking.md	ordered list	complete gene order
analysis/*-stats.json	structured	leader and mean (two checks)
analysis/*-profile.md	prose	top-three share
analysis/*-relations.md	prose	cross-chart rank
docs/dashboard.md	table	chart row in the shared dashboard
analysis/overview.md	prose	atlas-wide record holder
tests/test_atlas.py	executable	assertions over the current source
figs/<chart>.svg	vector graphic	painted bar count/order/relative lengths; numeric leader label
analysis/atlas-stats.md	prose	atlas-wide concentration

The scored downstream sites were fixed before the GPT-5.6-SOL evaluation.

D BixBench-RNA evidence and policy cases

The evidence/policy module contains five standalone workspaces and one compound project. The standalone cases isolate one transition; the compound case applies a sample correction, evidence withdrawal, and policy revision to shared findings, manuscript, figure, and validation queue.

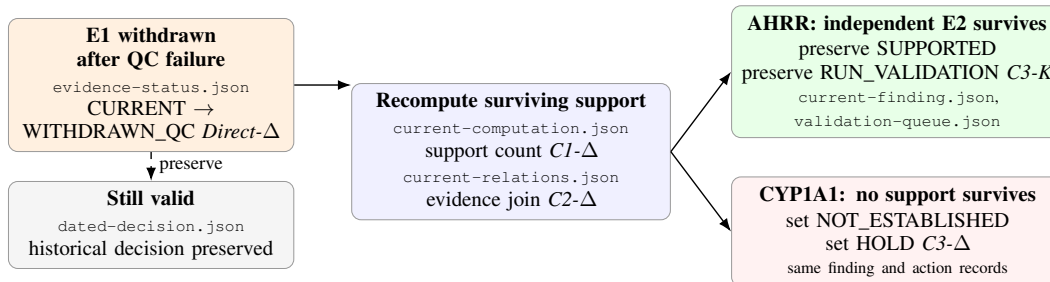


Figure 4: The same evidence withdrawal requires opposite conclusion operations. AHRR’s independent E2 support preserves its finding and action, whereas CYP1A1’s lost final support requires revising both. The three GPT-5.6 configurations perform this finding/action pair correctly.

E Ethics and AI use

The benchmark contains no private or human-subject biomedical data and is not intended for clinical use. Generative AI assisted code, experiments, and writing; the authors reviewed all reported claims and aggregates.

F Restoration and abstention details

Table 9: File-level outcomes for the restoration and abstention episodes summarized in Table 6. Temporary files retain the withdrawn correction; third-state files match neither the original nor the corrected version.

Transition	Agent	Scored file behavior	Episode success
Correction withdrawn	GPT-5.6-LUNA	822/1,035 restored; 202 third states; 11 temporary states	0/60
	GPT-5.6-TERRA	674/1,035 restored; 360 third states; 1 temporary state	0/60
	GPT-5.6-SOL	800/1,035 restored; 235 third states; 0 temporary states	0/60
	Claude Haiku 4.5	895/1,035 restored; 137 third states; 3 deleted; 897 collateral edits	1/60
	Claude Sonnet 5	848/1,035 restored; 184 third states; 3 temporary; 366 collateral edits	11/60
	Claude Opus 5	841/1,035 restored; 194 third states; 6 collateral edits	2/60
Already applied	GPT-5.6-LUNA	91/5,420 prose/data files and 4/870 figures changed	1/20
	GPT-5.6-TERRA	128/5,420 files and 69/870 figures changed	3/20
	GPT-5.6-SOL	16/5,420 files and 1/870 figures changed	10/20
	Claude Haiku 4.5	278/5,420 files and 25/870 figures changed	0/20
	Claude Sonnet 5	152/5,420 files and 0/870 figures changed	8/20
	Claude Opus 5	18/5,420 files and 0/870 figures changed	13/20