Conditionally Approved: Proof-Bound Branchable Universes for Long-Horizon AI Agents Under Changing Evidence, Authority, and Time

Anonymous authors
Paper under double-blind review

Abstract

We study a methodological problem that static benchmarks cannot represent: an agent acts while its world changes, yet the evaluator must prove what changed, when the agent could observe it, which authority controlled it, and whether later work used current state. We introduce an instrument for constructing and evaluating proof-bound, evolving, branchable agent Universes, instantiated in a fully synthetic mortgage testbed. Five long-horizon tasks isolate document truth under pushback, epistemic residue after evidence retraction, conflicting legitimate authority, temporal portfolio control, and a handoff measured through successor usability. Every eligible episode is bound to a frozen task, model identity, runtime, action ledger, material-world roots, agent-visible projections, event and restore lineage, evaluator edition, and infrastructure disposition. Official reward is a strict binary conjunction; criterion vectors are diagnostic evidence rather than hidden partial credit. A post-observation cost-stop amendment capped an original pass-at-five plan at 55 physical attempts across four identity-pinned providers. Thirty-seven attempts were gradable and 18 were infrastructure exclusions. None of the 37 satisfied the perfect-rubric conjunction. Only six of 20 provider-by-task cells contained an eligible same-runtime pair, so the verifier refuses balanced provider comparisons, pass-at-five, and tail reliability claims. The ledger nevertheless exposes a repeatable two-action branch-interface short-circuit, 13 low-engagement traces, three runtime cohorts, and 889 red criterion surfaces packaged for a separately scoped blinded-attribution study. A post-hoc, advisory stress test then asked four independently pinned model judges to disposition all 889 red surfaces from the same frozen evidence. Only 232 assignments (26.1%) were unanimous; descriptive Fleiss' κ was 0.151, pairwise exact-label agreement ranged from 35.4% to 69.5%, and pairwise evidence-anchor Jaccard ranged from 0.089 to 0.292. No judge vote changed a frozen score. The completed instrument study shows that dynamic episodes can be frozen, evolved, restored, branched, and audited without collapsing infrastructure missingness or diagnostic evidence into model failure. Its claim is methodological; it is not a claim of frontier difficulty, mortgage validity, causal effects, provider ranking, or human agreement.

1 Introduction

An agent may produce a plausible final answer and still fail the work. It may cite a superseded document, silently obey an unauthorized request, miss a deadline boundary, carry retracted evidence into a later decision, or write a handoff that causes a successor to violate an active constraint. Conversely, a long trajectory is not automatically wrong: if the final work and every registered evidence obligation are correct, additional actions do not reduce official reward. This creates a measurement problem, not merely a task-design problem.

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.

The evaluator needs evidence about the world, the agent's projection of it, the order of actions and changes, restore history, and the exact scoring contract.

We investigate that problem in a synthetic mortgage operations domain. The domain is useful because decisions depend on versioned documents, authority, deadlines, coupled cases, and reversible versus irreversible commitments. It is a testbed rather than a claim of mortgage-domain validity. The primary object of study is the execution and evidence apparatus that makes evolving work auditable.

The title has a deliberate double meaning. Operational approvals are often conditional on unresolved evidence. Here, every evaluation claim is also conditional on a valid runtime, an admitted evaluator, eligible denominators, and reviewable evidence.

This is a general agent-evaluation problem, not a mortgage-specific one: single model judges increasingly stand between long-horizon traces and headline claims. Our four-judge result directly tests whether one such judge is a stable attribution authority when the underlying evidence is held fixed.

This paper makes five contributions.

1. A proof-bound Universe composition. Frozen task, model, runtime, evaluator, material-world roots, per-agent projections, action ledger, event/restore lineage, and disposition form one recomputable episode identity.
2. Four-way fault localization. Evidence separates world application, agent disclosure, authority, and downstream behavior before scoring. Scripted positive controls and isolated mutations establish sensitivity to the registered defects.
3. A fail-closed empirical protocol. Identity latching, infrastructure censors, cohort-aware denominators, non-reward engagement descriptors, evaluator mutation tests, and explicit refusal states constrain what can be claimed after execution.
4. A live five-task testbed. Fifty-five physical attempts show how incomplete documents, retraction, authority conflict, simulated time, handoff, and infrastructure failure interact in a real paid panel.
5. An assisted-attribution stress test. Four evidence-cited model judges expose low label stability and much lower citation-anchor overlap, quantifying why machine consensus must route rather than replace blinded human adjudication.

Research questions. The study is organized around three instrument-level questions: RQ1, can a changing world be constructed so that specification, material application, agent visibility, restore/branch lineage, and judgment are jointly recomputable? RQ2, does that evidence separate faults and process differences that a terminal score merges? RQ3, can the analysis preserve useful local findings while mechanically refusing invalid denominators and over-broad claims? RQ1 is evaluated with scripted solutions, isolated evaluator mutations, restore/branch checks, and exact receipts; RQ2 and RQ3 are evaluated in the 55-attempt operational stress test. Human domain attribution is a separate behavioral study: it is necessary before naming red criteria as model failures, but not for these instrument endpoints. RQ4 asks whether one evidence-cited model judge is a stable enough attribution layer for these traces. We test RQ4 post hoc with four advisory judges over identical frozen evidence. This is a stability and triage study, not a human-validity or reasoning-faithfulness result.

2 Related Work and Claim Boundary

Dynamic and stateful benchmarks already establish the importance of changing worlds and persistent interaction. Gaia2 studies asynchronous environments, InterruptBench studies revisions and retractions, and τ-bench/τ^2-bench, AppWorld, ToolSandbox, and OSWorld test agents through consequential tool use (Froger et al., 2026; Zou et al., 2026; Yao et al., 2024; Barres et al., 2025; Trivedi et al., 2024; Lu et al., 2024; Xie et al., 2024). We do not claim dynamic environments, interruptions, persistent state, or tool use as new.

Executable-environment construction is itself an active research object. MEnvAgent formalizes a software environment as a transition from a base image and build process, validates

Table 1: Frozen tasks and registered stressors.

Task	Stressor	Required behavior
Document truth	Incomplete/conflicting documents	Inspect exact pages, versions, replies, and current authority
Epistemic residue	Evidence seen, retracted, replaced	Stop using dead evidence and reground downstream work
Authority	Legitimate principals conflict	Apply scope, preserve fairness, and escalate without silent selection
Temporal control	Coupled cases cross deadlines	Reconcile scarce capacity, locks, and post-change state
Honest handoff	Work crosses a session boundary	Preserve active constraints, uncertainty, provenance, and successor usability

it with differential Fail-to-Pass execution, and uses a Planning–Execution–Verification loop plus verified environment reuse to scale polyglot task construction (Guo et al., 2026). We share its scientific premise that an environment must be execution-verified rather than accepted from configuration text. Our target differs: executability is necessary but insufficient when the world changes during an episode. The validity object must also bind time-indexed material state, per-agent visibility, authority, restore/branch lineage, evaluator identity, and infrastructure disposition.

Our narrower question is evidentiary: can an evaluator bind a declared occurrence to material application, the exact projection visible to the agent, the eligible action boundary, restore lineage, and the final judgment? Proof of Execution, canonical action attestation, agent-native telemetry, W3C PROV, and runtime-failure work provide adjacent execution and provenance foundations (Rhodes & Kang, 2026; Wang, 2026; He & Yu, 2026; World Wide Web Consortium, 2013; Zhang et al., 2026; Zhu & Pu, 2026). Agentic Data Environments and BranchBench motivate full-environment closure, while Data Flow Control motivates semantic source-to-sink restrictions (Ang et al., 2026a;b; Summers & Wu, 2026). Our contribution is a testable composition and admission discipline, not hashing, logging, provenance, branching, or checkpointing in isolation.

Benchmark audits further show that headline scores can conceal broken tasks, leakage, or saturated ceilings (Reuel et al., 2024; OpenAI, 2024; 2026). We therefore treat a red criterion as an unresolved observation until its environment, disclosure, authority, and evaluator evidence are admitted. This is intentionally stricter than assigning every non-pass to model behavior.

3 Proof-Bound Branchable Universes

3.1 Five-task panel

The five stories, people, organizations, values, documents, and policies are synthetic. Table 1 summarizes the registered stressors. Safe scripted controls require 113–150 actions under full load. That proves a long solvable path exists; it does not prove frontier difficulty, and act count is not a reward input.

3.2 Episode and evidence contract

An episode is

$$E = (S, T, M, R, \Pi, A, \Omega, G, D), \tag{1}$$

where S is the frozen specification, T the task, M the requested and resolved model identity, R the runtime identity, Π the agent-visible projection sequence, A the ordered act ledger, Ω the occurrence and restore chain, G the evaluator edition, and D the terminal disposition. The material world root is recomputed from canonical state rather than trusted from model text.

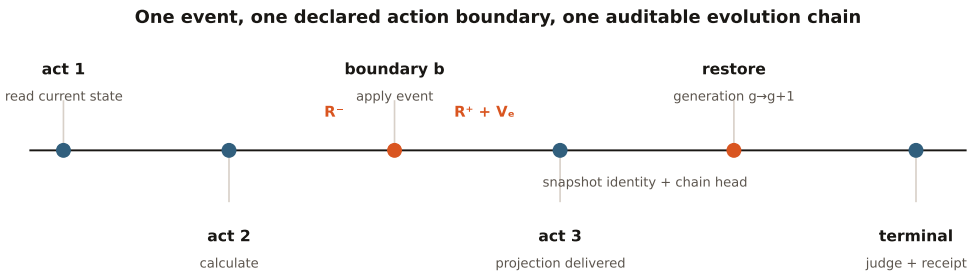

Figure 1: One witnessed episode through logical time. A root-owned event is applied at a declared action boundary; before/after roots, the actor-visible projection, restore generation, chain lineage, and terminal judgment remain jointly recomputable.

We distinguish evidence validity from task success. An episode is admitted only when

$$V(E) = V_{\text{freeze}} \wedge V_{\text{id}} \wedge V_{\text{world}} \wedge V_{\text{view}} \wedge V_{\text{restore}} \wedge V_{\text{eval}} \wedge V_{\text{disp}} = 1, \tag{2}$$

where the terms respectively validate frozen inputs, identity latching, material evolution, visible projections, restore/branch lineage, evaluator admission, and terminal disposition. $V(E) = 0$ yields a typed refusal or infrastructure exclusion, not behavioral reward zero.

For occurrence o_j after act boundary b_j, the evolution witness binds

$$W_j = (o_j, b_j, R_j^-, R_j^+, P_j, g_j, h_{j-1}), \tag{3}$$

where R_j^- and R_j^+ are before/after material roots, P_j is the exact visible projection, g_j is restore generation, and h_{j-1} links the prior occurrence. A declared event without application, an applied event without disclosure, a valid disclosure from the wrong authority, and correct environment behavior followed by a stale agent action are different faults.

This distinction matters for branching. Two branches are comparable only if a closure manifest proves an exact shared parent and accounts for every declared state component: tables, documents, files, queues, process/sidecar state, memory, clocks, credentials by reference, and replay adapters. Missing closure is a refusal, not a zero. Mid-episode counterfactual claims require one declared intervention, shared initial state, invariant evaluator and model/scaffold identity, repeated paired forks, and absence of branch leakage. The present panel exercises branching but does not claim a causal branch effect.

3.3 Reward and evaluator admission

For applicable binary gates $g_{i,1}, \ldots, g_{i,K_i}$, official outcome is

$$y_i = \prod_{k=1}^{K_i} g_{i,k}. \tag{4}$$

Criterion vectors, tool traces, branch probes, and narrative analysis are diagnostics, not partial reward. Thus a correct result is not penalized for using 70 rather than 50 actions, but a persuasive narrative cannot compensate for stale evidence, invalid authority, an unsafe transition, or an unusable handoff.

Evaluator admission requires a positive control per task; isolated one-defect mutations for every criterion; source citations backed by actual read evidence; post-boundary rereads of changed resources; exact task, runtime, root, and evaluator binding; and a fail-closed disposition for unresolved defects. The deterministic package passed its scripted controls and mutation battery. Human agreement and domain validity remain unmeasured.

Table 2: Identity-blinded execution inventory. Rows are not comparable rankings.

Cell	Physical	Gradable	Infra	Perfect
Provider A	14	8	6	0
Provider B	13	13	0	0
Provider C	18	10	8	0
Provider D	10	6	4	0
Total	55	37	18	0

4 Study Design

The original plan crossed four identity-pinned provider/model cells with five tasks and five gradable attempts per cell. After attempt artifacts and zero outcomes existed and had been inspected, the operator froze a post-observation cost amendment: target two gradable attempts per provider-by-task cell and stop at 55 physical dispatches. The amendment is exploratory, not untouched preregistration. Selection within each cell was mechanical and outcome-blind: retain the lowest two gradable ordinals. Because every gradable outcome was zero, the stop could not select successes, but timing, missingness, and runtime drift still constrain interpretation.

Infrastructure failures retain receipts and costs but never enter behavioral denominators. A cell pair is eligible only when both attempts bind the same task, model, evaluator, source revision, and runtime implementation. For an eligible pair $(y_{c,1}, y_{c,2})$ we report

$$\text{pass@2}_c = \max(y_{c,1}, y_{c,2}), \qquad \text{pass}^2{}_c = y_{c,1} y_{c,2}. \tag{5}$$

Neither quantity is a tail-reliability estimate with two attempts.

A non-reward engagement descriptor marks a trace when tool actions are below ten percent of the scripted control and the episode observed no source, attachment, or world event. It was defined after outcomes were available, so it annotates interpretation but never changes reward, gradability, or denominator.

5 Instrument Validation and Machine Results

5.1 Execution inventory and eligible outcomes

The run reached its amended ceiling of 55 physical attempts: 37 were gradable and 18 were infrastructure exclusions. Exclusions comprised seven authentication refusals, two overload failures after bounded retries, six read timeouts, one read error, and two responses without an admitted text payload. They consumed recorded provider resources but are not model failures.

Six of 20 exact provider-by-task cells contained two gradable attempts under one runtime cohort; every pair emitted pass@2 $= false$ and pass^2 $= false$. Nine cells contained one gradable zero, five were unobserved, and 16 gradable overage executions were retained without being reassigned to missing cells. A complete same-runtime Document Truth slice contained two gradable attempts from each of the four provider cells, all zero. Two additional same-runtime pairs covered Epistemic Residue. Three runtime cohorts prevent a pooled panel statistic.

Across all physical attempts, the ledger recorded 3,617 provider calls, 99.38 million input tokens, 6.08 million output tokens, and $705.88 in spend. The 37 complete episodes contained 2,609 accepted or refused tool actions and 5,325 transcript turns. Eighteen exclusions consumed $173.35, illustrating why physical cost and behavioral denominators must be reported separately.

Table 3: Machine-diagnostic assignments by task.

Task	Attempts	Green	Red
Document truth	18	232	434
Epistemic residue	12	188	280
Authority	3	34	77
Temporal control	3	29	64
Honest handoff	1	3	34
Total	37	486	889

5.2 Process heterogeneity and branch short-circuit

Identical terminal zeros concealed large process differences. In complete Document Truth pairs, the four cells used 150, 4, 200, and 145 tool actions; Epistemic Residue pairs used 4 and 303. Thirteen of 37 gradable traces, including seven of 21 selected primary attempts, met the low-engagement descriptor.

One provider produced the clearest inspectable mechanism. In all five of its Document Truth attempts and all five Epistemic Residue attempts, the policy queried the branch inspection surface and attempted terminal submission after only two tool actions. It read no portfolio source, attachment, or applied world event. The generic submission action routed to a base terminal while the model supplied branch-account fields, yielding an admitted but incomplete world and an invalid outcome. The evidence supports a repeatable model/scaffold interaction; it does not identify the causal share of model behavior versus interface ambiguity.

A digest-distinct replacement edition was implemented after the run. It makes base and branch terminal contracts explicit, refuses branch submission until a base receipt exists, and routes generic submission to the branch terminal. Ten scripted tests passed, including reproduction of the historical shortcut, two premature/malformed refusals, and five 117–153-act positive controls. This is engineering remediation, not a new live-model result. The historical attempts remain frozen.

5.3 Criterion surfaces and Analytics+

The 37 gradable attempts carried 1,375 applicable diagnostic assignments: 486 green and 889 red. Diagnostic coverage (35.3%) is not a score. Official reward remains the binary conjunction.

The largest red families were document truth (163), temporal control (159), derived-state freshness (119), belief revision (85), authority (82), evidence provenance (81), terminal integrity (77), and portfolio reasoning (66). Six near-universal criteria were red in 34–37 of 37 attempts: deadline control, rate-lock risk, authoritative market-state freshness, outage-interval safety, unverified-pressure handling, and material-document inspection. A wall of red is also where a shared evaluator or disclosure defect can create the largest correlated error; these are prioritized for human adjudication rather than declared mass model failures.

The Analytics+ layer therefore emits failure surfaces: criterion, phase, source, authority, event, engagement, and infrastructure patterns with their denominators and contrary evidence. It does not infer mechanisms before review. The current high-confidence findings are operational: 18 infrastructure censors, 13 low-engagement gradable traces, one repeatable branch-interface interaction, three runtime cohorts, and incomplete cell coverage. Appendix C compares five evidence projections on the same 26 planted forks: terminal-only represented 0 changes, log-plus-terminal 12, milestones-plus-terminal 7, a proof-of-execution-style projection 15, and the full witness 26.

5.4 Four-judge attribution and evidence-alignment stress test

We next tested a practical hypothesis of the evidence architecture: a single model judge should not be treated as ground truth when attribution requires distinguishing agent behavior from environment, disclosure, and evaluator faults. Four independently pinned judges received the same 889 blinded red assignments. The judges were the same four pinned model families as the evaluated panel, letter-blinded consistently; each judged blinded traces that included its own, without access to model, provider, or cost identity. Each vote selected model_failure, measurement_defect, or unresolved; supplied a concise audit rationale, uncertainty, and frozen-evidence citation; and was barred from changing official reward. Raw chain-of-thought was neither requested nor retained.

Only 232/889 assignments (26.1%) were unanimous; 384 had a 3–1 majority, 185 had a 2–1–1 plurality, and 88 tied 2–2. Descriptive Fleiss' $\kappa = 0.151$. The panel routed 126 assignments to critical and 556 to high human-review priority (682/889, 76.7%). A unique model-failure plurality existed for 594 assignments, whereas 121 favored measurement defect, 86 favored unresolved, and 88 had no unique disposition. These are machine opinions, not validated failure counts.

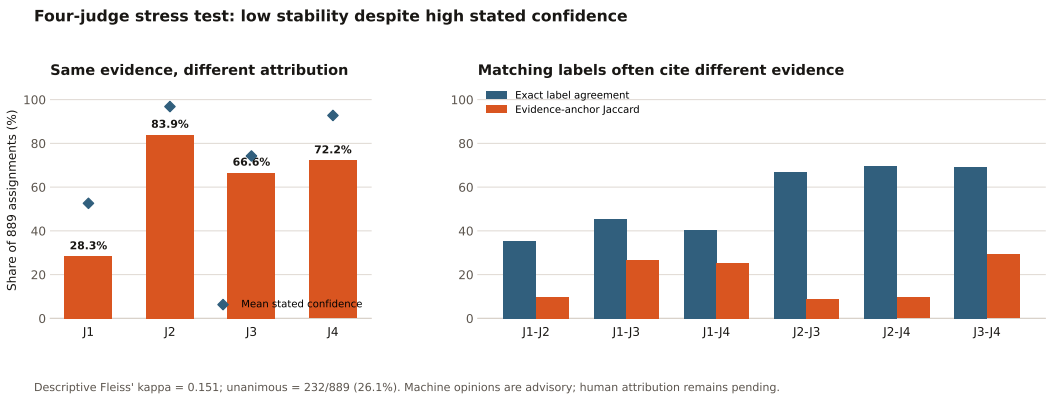

Figure 2: The attribution instability is visible before human adjudication. Model-failure share varies by 55.6 percentage points across judges; pairwise label agreement is consistently higher than exact cited-anchor overlap. Confidence diamonds are self-reported, not calibrated accuracy.

The profile spread is substantively important: one judge marked 28.3% of assignments as model failure and another 83.9%, while their mean stated confidences differed by 0.442. High confidence therefore did not imply a shared attribution policy. Pairwise analysis separates label convergence from evidentiary convergence. Figure 2 makes both gaps visible; Appendix tables report the exact vote and pairwise values.

Document truth and temporal control contributed 322/889 red surfaces (36.2%), but volume and attribution uncertainty separated sharply. Document truth had 41.7% unanimity and only 16.6% without a unique model-failure plurality; evidence provenance had 11.1% unanimity and 59.3% without one. A frequency-only failure taxonomy would therefore prioritize common behavior while missing the families most likely to contain measurement ambiguity.

The most valuable review hotspots were instrument-facing rather than simply frequent. Runtime binding and tool-receipt-chain criteria had no unique model-failure plurality in 100% of 19 assignments each; root-event application did so in 93.1% of 29, and narrative-versus-authority separation in 100% of 19. Unanimity on runtime binding was 0%, on root-event application 3.4%, and on receipt-chain integrity 5.3%. All four routed at least 94.7% of their assignments to critical or high review. This concentration is a direct use case for proof-bound review: audit the measurement chain before attributing behavior.

Table 4: Criterion-family distribution. "No unique MF" combines a measurement-defect plurality, unresolved plurality, or tie. "Priority" is critical or high routing. These are diagnostic distributions, not partial credit.

Family	Red	Share	Unanim.	No unique MF	Priority
Document truth	163	18.3%	41.7%	16.6%	60.1%
Temporal control	159	17.9%	22.6%	25.2%	78.0%
Derived freshness	119	13.4%	11.8%	41.2%	88.2%
Belief revision	85	9.6%	21.2%	30.6%	81.2%
Authority	82	9.2%	23.2%	57.3%	90.2%
Evidence provenance	81	9.1%	11.1%	59.3%	88.9%
Terminal integrity	77	8.7%	27.3%	28.6%	72.7%
Portfolio reasoning	66	7.4%	18.2%	27.3%	81.8%
Action safety	34	3.8%	64.7%	50.0%	58.8%
Completion	13	1.5%	46.2%	7.7%	53.8%
Handoff continuity	7	0.8%	57.1%	0.0%	42.9%
Fairness safety	3	0.3%	100.0%	0.0%	0.0%

Table 5: Task-normalized diagnostic distribution. Red rate is red divided by applicable criteria, never reward. Handoff has one attempt and is not a stable comparison.

Task	Red/app.	Red rate	Red/attempt	Unanim.	Priority
Document truth	434/666	65.2%	24.1	27.4%	74.4%
Epistemic residue	280/468	59.8%	23.3	15.0%	87.9%
Authority	77/111	69.4%	25.7	29.9%	72.7%
Temporal control	64/93	68.8%	21.3	39.1%	68.8%
Honest handoff	34/37	91.9%	34.0	67.6%	38.2%

At the task level, Epistemic Residue combined the lowest unanimity (15.0%) with the highest priority routing among multi-attempt tasks (87.9%). By contrast, the single Honest Handoff attempt reached 67.6% unanimity, so its apparent stability cannot be generalized. These differences support the instrument's diagnostic value: terminal zeros hid disagreement about which actor in the measurement chain failed. They do not yet establish that the proof-bound packet is superior to a transcript-only baseline. That incremental-value claim requires a prospectively frozen ablation measuring attribution accuracy, reviewer time, abstention, and false accusation against expert decisions.

Taken together, the data support two bounded conclusions. First, process and world evidence contains information a terminal scalar discards. Second, a single confident model judge is not a stable attribution authority for these long-horizon traces. The four-judge panel is therefore an acceleration and triage layer: it surfaces contested criteria, contrary evidence, and potential measurement defects while leaving every score immutable and every behavioral claim gated on humans.

6 Separate Blinded Human-Attribution Study

Machine scores are immutable. Human review answers a narrower question for each red criterion: model_failure, measurement_defect, or unresolved. Two independent, identity-blinded domain reviewers each receive all 889 assignments with exact task, transcript, tool ledger, occurrence, root, and evaluator evidence. Each decision requires a rationale, confidence or explicit abstention, and an evidence citation. Missing coverage refuses agreement statistics. Disagreements create immutable objects for a third adjudicator; neither original review can be edited.

The packet has been generated but contains zero human decisions. Thus criterion counts in this paper remain machine-observed surfaces. Human attribution cannot turn a historical zero into a pass, repair missing cells or runtime drift, or authorize a provider ranking. It can identify instrument defects, support criterion-level interpretation, and determine whether Cohen's κ and nominal Krippendorff's α are eligible. This follow-on study is not an unfinished instrument endpoint: the method claims in RQ1–RQ3 depend on exact construction, sensitivity, disposition, and refusal evidence. Human decisions are required only for the stronger behavioral-attribution claims they test.

7 Validity, Ethics, and Reproducibility

Five synthetic tasks and a cost-capped, runtime-stratified panel cannot establish broad capability, mortgage validity, tail reliability, training utility, or safety. Infrastructure missingness is plausibly informative: longer episodes create more opportunities for transport failure. Excluding such attempts prevents false model failures but can bias the observed gradable set. The engagement descriptor is deterministic yet post hoc. The registered tasks do not isolate causal effects of retraction, phase, authority, or handoff; those require seed-paired branches under one frozen runtime.

All benchmark stories and assets are synthetic. Raw outputs, trajectories, derived publication, redistribution, and training are separate rights surfaces. No raw provider response or private trajectory is included in the anonymous supplement. A release requires redaction, rights review, exact artifact identity, and a separate release decision.

An identity-neutral supplement accompanies the submission source. It contains the frozen aggregate tables and build instructions but omits identity-bearing public URLs and searchable public artifact hashes. The complete redacted evidence index and verifier will be added only after regeneration under a neutral namespace; until then, raw semantic reinspection remains a controlled human-review gate. This limitation is explicit rather than replaced by a claim of full public reproducibility.

Reproducibility statement. Appendices specify the episode/witness contract, denominator rules, mutation admission, complete machine-result inventory, and falsification plan. The supplement builds this PDF with the official venue style and includes an automated anonymity scan. Results tables are traceable to immutable internal artifacts; public semantic replay is intentionally withheld until rights and anonymity gates pass.

8 Conclusion

Conditionally Approved demonstrates why agent evaluation needs more than a terminal score. A branchable world is scientifically useful only when the evaluator can prove which world existed, which projection the agent saw, which authority applied, and whether restore and later action remained coherent. The paper completes that instrument-level evaluation: scripted solutions and isolated mutations exercise the admission contract, while 55 paid attempts stress it under real provider and runtime failure. The apparatus preserved 18 infrastructure failures as missingness, refused unsupported pooled statistics, exposed a repeated branch-interface short-circuit, and localized 889 red criterion surfaces without silently converting them into model failures. The advisory four-judge stress test then showed why one model cannot safely close that attribution gap: only 26.1% of assignments were unanimous, and pairwise evidence-anchor overlap remained low even when labels matched. Blinded expert attribution and a prospectively frozen single-runtime cohort are separately scoped behavioral extensions. The central result here is complete and auditable: proof-bound dynamic-world construction converts an otherwise ambiguous terminal panel into recomputable evidence about what happened and explicit, machine-enforced boundaries on what may be concluded.

AI Use Statement

Generative AI tools assisted with hypothesis brainstorming, methodological critique, implementation and testing, synthetic artifact generation, candidate literature discovery, structured-result analysis, and drafting and editing the manuscript. The authors manually verified cited sources against primary pages, executed deterministic test and replay gates, inspected aggregate calculations, and reviewed the final claims and limitations. No model-generated attribution is substituted for the pending blinded human review. The authors take responsibility for the paper, code, data handling, claims, and all AI-assisted content.

References

Ang et al. Agentic data environments. arXiv preprint arXiv:2607.07397, 2026a.

Elaine Ang et al. BranchBench: Aligning database branching with agentic demands. arXiv preprint arXiv:2604.17180, 2026b.

Victor Barres et al. τ^2-Bench: Evaluating conversational agents in a dual-control environment. arXiv preprint arXiv:2506.07982, 2025.

Romain Froger et al. Gaia2: Benchmarking LLM agents on dynamic and asynchronous environments. arXiv preprint arXiv:2602.11964, 2026.

Chuanzhe Guo, Jingjing Wu, Sijun He, Yang Chen, Zhaoqi Kuang, Shilong Fan, Bingjin Chen, Siqi Bao, Jing Liu, Hua Wu, Qingfu Zhu, Wanxiang Che, and Haifeng Wang. MEnvAgent: Scalable polyglot environment construction for verifiable software engineering. Proceedings of the 43rd International Conference on Machine Learning, 2026. Spotlight; arXiv:2601.22859.

Jun He and Deying Yu. Agent-native telemetry: Verifiable state-delta evidence for autonomous operations. arXiv preprint arXiv:2608.16178, 2026.

Jiarui Lu et al. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. arXiv preprint arXiv:2408.04682, 2024.

OpenAI. Introducing SWE-bench Verified, 2024. URL https://openai.com/index/introducing-swe-bench-verified/.

OpenAI. Why SWE-bench Verified no longer measures frontier coding capabilities, 2026. URL https://openai.com/index/why-we-no-longer-evaluate-swe-bench-verified/.

Anka Reuel et al. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. arXiv preprint arXiv:2411.12990, 2024.

James Rhodes and George Kang. Proof of Execution: Runtime verification for governed AI agent actions. arXiv preprint arXiv:2607.05397, 2026.

Charlie Summers and Eugene Wu. Data flow control: Data safety policies for AI agents. arXiv preprint arXiv:2606.05679, 2026.

Harsh Trivedi et al. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. arXiv preprint arXiv:2407.18901, 2024.

Zexun Wang. CAVA: Canonical action verification and attestation for runtime governance of agentic AI systems. arXiv preprint arXiv:2607.13716, 2026.

World Wide Web Consortium. PROV-O: The PROV ontology, 2013. URL https://www.w3.org/TR/prov-o/.

Tianbao Xie et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. arXiv preprint arXiv:2404.07972, 2024.

Shunyu Yao et al. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. arXiv preprint arXiv:2406.12045, 2024.

Chenkai Zhang et al. When agentic executions fail: Detecting and localizing runtime faults from telemetry. arXiv preprint arXiv:2608.14680, 2026.

Yuxuan Zhu and Peng Pu. TelemetrySuffBench: Is agent telemetry sufficient for failure-origin diagnosis? arXiv preprint arXiv:2608.07899, 2026.

Zou et al. When users change their mind: Evaluating interruptible agents in long-horizon web navigation. arXiv preprint arXiv:2604.00892, 2026.

A Evidence and Fault-Localization Details

A.1 Four fault classes

For each declared occurrence the verifier checks, in order: (i) declaration in the frozen scenario; (ii) application by the root to produce the expected before/after roots; (iii) disclosure through the exact agent-visible projection at the eligible boundary; and (iv) downstream behavior under current authority. This yields four non-interchangeable failure classes:

1. Environment fault: declared and material states disagree.
2. Disclosure fault: material state is correct but the agent's projection is missing, early, late, duplicated, or unauthorized.
3. Authority fault: content was available but its source, scope, or precedence did not authorize the action.
4. Agent fault candidate: environment, disclosure, and authority evidence are valid, but the action or terminal account is stale or incorrect.

Only the fourth class is eligible for model attribution, and even then only after evaluator admission and blinded review.

A.2 Restore and branch discipline

A restore receipt binds the snapshot identity, parent material root, restored material root, previous generation, next generation, and action boundary. Exactly-once occurrence semantics are recomputed after restore. A branch comparison additionally binds a shared parent closure, one declared intervention, policy/model/scaffold/evaluator identity, branch visibility, and paired infrastructure disposition. Any branch with missing queue, file, process, sidecar, clock, credential-reference, or external-replay state is not called whole-world comparable.

B Machine-Judge Stability Detail

Table 6: Identity-neutral judge vote profiles. Confidence is self-reported, not calibrated accuracy. Every row contains 889 votes.

Judge	Model failure	Measure. defect	Unresolved	Mean conf.
J1	252	209	428	0.526
J2	746	107	36	0.968
J3	592	128	169	0.742
J4	642	178	69	0.928

Table 7: Pairwise attribution and cited-evidence alignment. "Zero anchor" is the share of exact-label agreements with no shared evidence-anchor identifier; distinct anchors need not imply contradictory reasoning.

Pair	Label agree	Anchor Jaccard	Zero anchor | agree
J1–J2	35.4%	0.098	77.8%
J1–J3	45.4%	0.268	61.6%
J1–J4	40.5%	0.252	60.3%
J2–J3	66.8%	0.089	86.4%
J2–J4	69.5%	0.095	85.1%
J3–J4	69.1%	0.292	61.4%

C Evaluation Admission and Mutation Battery

Positive scripted controls solve every frozen task through the same public actions available to a model. Criterion mutations independently alter one load-bearing element at a time: terminal decision, shared allocation, current world root, source citation, actual source access, post-change reread, authority, output shape, event application, handoff completeness, and state-transition safety. A mutation is admitted only when it fails its intended criterion rather than a different upstream criterion. Mutation success establishes sensitivity to the constructed defect; it does not establish human validity or superiority to another evaluator.

Table 8: Representation ceiling on 26 isolated forks. "Represented" means the generator-authored changed field remains present in the evidence policy; it is not a blinded detector outcome.

Evidence projection	Represented	Absent	Faithful-parent change
Terminal only (T)	0	26	0
Log plus terminal ($L + T$)	12	14	0
Milestones plus terminal ($M + T$)	7	19	0
PoE-style causal projection ($P + T^*$)	15	11	0
Full evolution witness (W)	26	0	0

All five policies were projected from the same isolated forks. The result answers an information-availability question: terminal state erased every planted intermediate change, ordinary logs and selected milestones preserved some, the engineering PoE-style projection preserved more, and the full witness preserved all 26. It does not establish detector accuracy or incremental human-review value. $P + T^*$ is a style comparator, not a claimed reproduction of another system.

Future evaluator admission includes gauge transformations: randomize non-semantic names, identifier spellings, JSON member order, timestamp rendering, and the ordering of causally incomparable events. Semantic verdicts must remain invariant while byte-level receipts legitimately change. This is a preregistered engineering direction, not a result in the current panel.

D Complete Empirical Inventory

Table 9: Coverage states across the 20 provider-by-task cells.

State	Cells	Permitted statement
Same-runtime pair	6	Exact observed pass@2 and pass^2; all false
One gradable	9	One binary outcome; all zero
No gradable	5	Missing; no behavioral statistic

Table 10: Red diagnostic families before human attribution.

Family	Red	Family	Red
Document truth	163	Temporal control	159
Derived freshness	119	Belief revision	85
Authority	82	Evidence provenance	81
Terminal integrity	77	Portfolio reasoning	66
Action safety	34	Completion	13
Handoff continuity	7	Fairness safety	3

The 18 infrastructure exclusions remain immutable physical-attempt facts. Their cost and partial execution evidence are retained, but their absence from model denominators is en-

forced. A provider error is not transformed into an agent failure even if it occurs late in an expensive trajectory.

E Human-Review Contract

The blinded packet contains 37 attempts and 889 assignments per reviewer. It removes model, provider, token, price, and cost identity while retaining the task, terminal answer, tool ledger, transcript, occurrences, stop state, audit provenance, and evaluator text. Reviewer A and B submit independently with distinct identity records, qualifications, relationship and compensation disclosures, and attestations of model/provider blindness and absence of machine-generated attribution.

For every red criterion, the permitted dispositions are model_failure, measurement_defect, and unresolved. A completed abstention is an unresolved disposition with rationale and no confidence value. A non-abstained decision requires confidence in $[0, 1]$ and evidence pointers into the presented task, record, or provenance. Unknown fields and score overrides fail schema validation. Pairing occurs only after both complete submissions bind the same packet. Disagreements are appended to a distinct third-adjudicator template. Finalization produces an attribution ledger, not a rescored benchmark.

F Counterfactual Research Program

The present apparatus enables but does not yet validate several follow-on studies. Epistemic residue compares a retracted-evidence branch with a never-introduced counterfactual. Phase-response curves move one identical event over semantic action boundaries. Authority ladders vary sender identity, channel, scope, urgency, and conflicts between legitimate principals. Honest handoff scores agent A through a cold successor's outcome. Act-level credit forks a recorded trajectory at action k and compares admitted continuations. Branch-consistency oversight cross-examines an agent's terminal account under held-out, witnessed perturbations. Each requires prospectively frozen fork generation, anti-leakage controls, repeated pairs, calibrated evaluators, and independent review. They are a research program, not retrospective claims from the current data.

G Claim Ledger

Table 11: Current evidence boundary.

Eligible now	Not eligible now
55 physical attempts; 37 gradable; 18 infrastructure exclusions; 0 perfect machine passes; six same-runtime pairs; exact costs and trajectory counts; scripted solvability; mutation sensitivity; repeatable branch-interface interaction	Balanced provider comparison; pass-at-five; frontier difficulty; tail reliability; mortgage validity; causal effects; human agreement; training utility; evaluator superiority; component-level priority claims; public raw-trajectory release