



American Journal of Financial Technology and Innovation (AJFTI)

ISSN: 2996-0975 (ONLINE)

VOLUME 4 ISSUE 1 (2026)



PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

Stateful Mediation and Selective Auditing for Financial Language-Model Actions: A Paired Controlled Simulator Study

Ayush Ojha^{*}

Article Information

Received: June 29, 2026

Accepted: September 08, 2026

Published: September 30, 2026

Keywords

*Condition-Blind Proposal,
Financial Agent, Post-State
Harm, Selective Audit, Stateful
Mediation*

ABSTRACT

Language-model agents can propose financial actions based on observations that become stale before execution. This study measures how five execution arrangements translate the same model proposals into post-state harm and correct completion in a controlled synthetic financial workflow. A protocol was internally frozen before generation. Four fixed quantized local model builds and two seeds converted 240 visible contexts into 1,920 context-level proposals. Mapping those proposals to 480 paired episode variants produced 3,840 proposal-by-episode units; replay through five arms yielded 19,200 transitions. State-bound checking produced 24 harmful transitions among 3,840 units, compared with 1,126 under direct execution (paired risk difference -0.2870 ; separate one-sided 95% generator-sensitivity bounds -0.2901 to -0.2836). On hazardous interpositions, its paired difference from snapshot checking was -0.8167 ; on paired benign interpositions, its correct-completion difference from coarse full-state checking was $+0.8167$. Those contrasts are algebraic sign mirrors over shared contexts, not independent confirmations. State-bound checking retained 81.67% eligible completion but did not eliminate harms visible at observation. All three contrast directions were structurally constrained by the nested mediators; the resampling bounds address nonzero magnitude under the synthetic generator, not discovery of an empirically contestable sign. In 60,000 selective-audit trajectories, no policy crossed either prespecified stopping threshold by 480 observations; all threshold times were censored at 481. Binding execution to current control state changed consequences in this fixed suite, while the conservative audit procedure did not yield an operational stopping certificate.

INTRODUCTION

Language-model agents increasingly connect text generation to operations that change external state. In a financial workflow, an apparently plausible answer can become a payment, refund, beneficiary update, disclosure, or ledger entry. That change in consequence means that answer quality alone is an incomplete endpoint. A proposal may be well-formed and consistent with an earlier snapshot while becoming unauthorized or unsafe before it reaches the commit boundary. Conversely, a control that blocks every changed state may prevent harm but also reject an otherwise valid action after an irrelevant metadata update. The central measurement problem is therefore not simply whether a model selected an expected label. It is how a fixed proposal interacts with changing state and an execution-time mediator to produce a verifiable post-state outcome.

Existing agent benchmarks have made tool use, long-horizon behavior, prompt injection, and stateful environments measurable. ReAct established a common reasoning-and-acting pattern for language models (Yao *et al.*, 2023). ToolEmu and AgentDojo showed how emulated tools and dynamic environments can expose harmful actions and prompt-injection failures (Debenedetti *et al.*, 2024; Ruan *et al.*, 2024). More recent systems enforce policies, certificates, or authorization at

the execution boundary. These systems make clear that commit-time checks, compare-and-swap, deterministic verification, and action gating are established mechanisms rather than contributions that can be claimed merely by combining them with a language model.

Financial action evaluation adds two practical difficulties. First, an intervention should preserve the model-visible request while varying hidden state conditions; otherwise a measured difference may reflect a different prompt rather than a different execution condition. Second, observed failures may be selectively audited. If high-materiality cases receive more review, an unweighted failure rate among audited cases does not identify the full-stream rate. Risk-limiting financial audits and inverse-probability methods provide relevant statistical foundations (Horvitz & Thompson, 1952; Shekhar *et al.*, 2023), while confidence-sequence methods support time-uniform uncertainty under stated assumptions (Howard *et al.*, 2021). Those methods still require a defined target, known positive review probabilities, and outcome labels that correspond to the claimed harm.

This article addresses a deliberately narrow question: in a balanced controlled simulator, how do five execution arrangements convert condition-blind local-model proposals into policy-blind post-state harms and correct completions, and what descriptive behavior do selective-

¹ Student, Georgia Institute of Technology, North Avenue Atlanta, 595 Pacific Ave, San Francisco, California, United States

^{*} Corresponding author's e-mail: aojha47@gatech.edu

audit policies show on the resulting consequence streams? The design fixes the visible proposal context across stable, hazardous-interposition, and benign-interposition variants. It then applies the hidden interposition after proposal generation and before deterministic mediation. The same proposal is replayed through every compatible arm. This paired design separates proposal generation from execution-time state handling within the simulator.

The study has three confirmatory hypotheses. H1 compares harmful-state-transition incidence under control-state-bound checking with direct execution across the complete tested suite. H2 makes the corresponding comparison with snapshot checking on hazardous interpositions. H3 compares correct completion with coarse full-state compare-and-swap on benign interpositions. The hypotheses use a fixed gatekeeping order and stratified pair-cluster resampling. Selective-audit results are secondary: four static policies receive matched expected budgets and are summarized descriptively, while an outcome-history adaptive policy is exploratory and is not treated as an equal-budget comparator.

The residual contribution is bounded empirical integration and paired consequence measurement in this frozen synthetic suite, not a new authorization primitive. The study joins fixed local-model proposals, paired visible contexts and post-proposal interpositions, five known mediation patterns, a policy-blind postcondition labeler, case-clustered contrasts, and selective-audit replay in one prospectively frozen simulator protocol. It reports null-relevant details as well as favorable contrasts: 24 harms remain under state-bound checking; H2 and H3 are deterministic sign mirrors; no audit policy reaches the prespecified stopping thresholds; materiality is synthetic; and the evidence does not validate deployment, compliance, or a population of models or financial institutions.

LITERATURE REVIEW

Financial-language-model evaluation also includes information-retrieval tasks. Mohammed and Lund (2026) assess a financial-document chatbot using context recall, faithfulness, context utilization, and context precision. Such measures concern the retrieved context and generated response; this study asks the additional question of what happens when a proposal is committed against changing financial state.

Tool-using language-model research distinguishes a model's generation process from the environment-facing actions that follow it. ReAct interleaves reasoning traces and actions, creating a general pattern for agents that call tools (Yao *et al.*, 2023). Agentic-model surveys describe planning, memory, tool use, and multi-agent coordination as recurring components, but also note that evaluations differ widely in task definitions and success criteria (Plaat *et al.*, 2025). Safety-oriented benchmarks then move beyond answer matching.

ToolEmu evaluates risky agent behavior in an emulated tool environment, AgentDojo uses a dynamic, stateful environment to study prompt injection and defenses, and SafeToolBench prospectively assesses whether models recognize instruction, tool, and joint risks before execution (Debenedetti *et al.*, 2024; Ruan *et al.*, 2024; Xia *et al.*, 2025). SafeToolBench includes financial tool-use risks, but it does not replay a fixed proposal across post-observation state changes and deterministic mediators. These lines motivate environment-state scoring, but they do not make a financial post-state label automatically valid; the state machine, authorization facts, and harm rules still have to be declared.

Several 2026 systems are especially close to the present study. Commit-Time Authorization defines a durable-effect boundary at which earlier authority evidence must remain fresh, causally prior, effect-bound, and eligible; its controlled invalidation suite shows that visible endpoint success and authorized completion can diverge (Santos-Grueiro, 2026). CapLease focuses on semantic replay and argues that identifier-local single-use tokens are insufficient without durable authorization state and transactional issue-prepare-commit transitions (Xu *et al.*, 2026). The Authority Frontier and Actuarial Action Interface price side-effect-bearing actions against reserve budgets and expose how much autonomy is released at different capital levels (H.-H. Chen, 2026). The Organizational Control Layer separates proposal generation from execution, applying role, policy, and economic constraints before actions reach the environment (Shi *et al.*, 2026). AID-Guard extends this boundary across provider commit, ambiguous delivery, and recovery by revalidating the approved request and current provider state, retaining one reservation until terminal or certified-no-effect resolution, and evaluating strict and typed authority profiles (Tong *et al.*, 2026). Its provider-effect lifecycle study does not replay condition-blind financial proposals through the five mediators or selective-audit design evaluated here.

Two other 2026 systems clarify the evaluation side. ClawMark evaluates multi-turn, multi-day coworker agents in stateful services whose environments change between turns, with deterministic post-state checkers rather than a language-model judge (Meng *et al.*, 2026). Vera generates executable safety cases, runs agents in isolated environments, and verifies outcomes from tool evidence and environment state (Feng *et al.*, 2026). Both demonstrate the value of changing-state environments and deterministic verification at larger scope than a single structured-decision benchmark.

Runtime enforcement systems further reinforce that boundary. AgentSpec provides customizable runtime enforcement for language-model agents (Wang *et al.*, 2025), while ToolSafe studies proactive step-level guardrails and feedback for tool invocation (Mou *et al.*, 2026). These systems differ in policy representation and recovery behavior, but each treats generation and execution control as separable. The five arms in the

present study should be read as controlled comparators drawn from this broader design space. The nondeployable authorization reference supplies a simulator ceiling by constructing the authorized requested effect directly; it is not a model, a practical baseline, or evidence that any production monitor can achieve the same result.

August 2026 work makes current-state authority and stateful governance especially explicit. SAGE-Fin contracts financial-market-agent authority under current market, account, policy, and dialogue state and rechecks prior authorization after state changes (Tang *et al.*, 2026). MasuGate defines policy-state serializability and evaluates stale authorization over shared budgets and inventory against request-context baselines (Peng & Wu, 2026). SAGE-Fin is a conformance and field-corroboration architecture, while MasuGate is a general concurrent-governance design evaluated in an LLM-free procurement workflow. Neither study replays one condition-blind financial-language-model proposal through the five mediators used here or couples that replay to the prespecified post-state harm, completion, and selective-audit trajectories.

Financial-agent evaluations have also argued that generic task success can hide domain-specific risk. Risk-first work on financial applications emphasizes that an agent can complete a task while violating constraints important to a financial workflow (Z. Chen *et al.*, 2025). Workflow-fidelity evaluation similarly asks whether the path and controls used by an agent conform to an intended payment process rather than judging only the final answer (Huang *et al.*, 2026). These observations motivate the current post-state endpoints, but they do not validate the particular synthetic workflows or materiality weights used here. No institution supplied cases, and no financial-domain panel adjudicated the labels.

Recent finance-specific neighbors further narrow the claim. FinToolBench evaluates runnable financial-tool use over 760 tools and 295 tool-required queries; its focus is tool retrieval and execution dimensions, not paired hidden-state interventions on a fixed proposal (J. Lu *et al.*, 2026). FinSafetyBench is a bilingual refusal benchmark for financial-crime and ethics requests under three attack settings, not an effectful transaction-state benchmark (Hou *et al.*, 2026). “From Tasks to Teams” audits multi-agent finance with ten probes across model, workflow, interaction, and system layers, whereas the present unit is a single proposal replayed through deterministic mediators (Z. Chen *et al.*, 2026). The agentic-commerce systematization organizes transaction-authorization and other cross-layer threats but does not conduct this paired mediator experiment (Mao *et al.*, 2026).

FraudBench and ReguSim already cover finance-specific mutable-state and action-enforcement evaluation. FraudBench places banking agents in shared mutable account state and includes history-dependent cases in which a later request can appear locally valid while being unsafe in sequence (Pai & Xian, 2026). Its conversational and adaptive-fraud scenarios use history-dependent

judging; they do not byte-match a condition-blind proposal and replay it through deterministic mediation arms. ReguSim separates stated reasoning, attempted action, deterministic execution enforcement, pre- and post-trade state and ledger updates, and monitor evidence in financial compliance evaluation (Y. Luo *et al.*, 2026). It studies regime, persona, incentive, and monitoring effects over trader decisions rather than paired post-observation interpositions of the same proposal or mediator-specific safety and completion contrasts.

Two additional 2026 records make that boundary still sharper. Li and Zhu (2026) audit benign self-evolution in simulated e-banking and report that capability gains can coincide with greater exposure to injected content and unauthorized state change; they also show that an execution-interface mismatch can suppress tool use and create misleading apparent safety. Their intervention varies learned agent artifacts under matched acquisition trajectories and sealed endpoints, whereas this study freezes each proposal and replays it through post-proposal hidden-state interpositions and deterministic mediators. FinVault's archived version 1 describes 31 regulatory case-driven sandbox scenarios with state-writable databases, explicit compliance and permission controls, audit logs, 856 adversarial cases, and 107 benign cases (Yang *et al.*, 2026). It judges vulnerability triggers from observable tool actions, business state, and interaction history and reports attack-success and false-positive rates. It does not report byte-matched condition-blind proposals, a paired post-proposal hidden-state interposition, five alternative mediator arms, a shared policy-blind harm and eligible-completion endpoint, paired cluster inference, or selective-audit replay. The current version 2 record is withdrawn; this comparison uses the still-accessible version 1 full text and does not infer misconduct from the withdrawal.

Adjacent control research also establishes typed deterministic procedures and stale-evidence evaluation. STAGE translates policy-scoped context into reviewed executable graphs with typed node results and a deterministic coordinator, including evaluation on a proprietary banking-dispute benchmark (M. Luo *et al.*, 2026). Its endpoints concern workflow task success, procedural depth, and repeated-run reliability, not proposal-to-commit staleness or paired harm and completion under alternative mediators. WirelessOpsAgent controls stale or inconsistent telemetry and execution-state decision episodes in wireless operations (Z. Lu *et al.*, 2026). It supplies a cross-domain current-evidence and repair benchmark, but not a financial fixed-proposal mediator replay or selective audit.

Taken together, these works rule out broad novelty claims about financial tool use, financial-language-model safety, stateful financial-action evaluation, current-state authorization, deterministic execution gating, post-state checking, typed control, or agentic-commerce threat modeling. Among primary sources

audited through 3 September 2026, the review identified no study combining all prespecified design elements: byte-matched condition-blind proposals, paired hidden-state interpositions, five deterministic mediation comparators, policy-blind harm and eligible-completion endpoints, pair-clustered contrasts, and prespecified selective-audit replay. The residual contribution is the bounded paired consequence measurement in this frozen synthetic suite, together with its prespecified descriptive audit replay. This statement is a dated search boundary, not a universal priority claim.

In the broader auditing literature, Raidah *et al.* (2026) discuss continuous transaction monitoring and full-dataset analysis while retaining a role for human oversight. The selective-audit component here estimates a full-stream synthetic material-risk target from sampled outcome labels under known review probabilities. This distinguishes transaction monitoring from complete access to verified outcome labels.

The selective-audit component draws on sampling and sequential inference rather than agent-control novelty. Horvitz-Thompson weighting corrects unequal inclusion probabilities when those probabilities are known (Horvitz & Thompson, 1952). Risk-limiting financial audits demonstrate how weighted sampling can focus review while retaining a declared error guarantee for a financial target (Shekhar *et al.*, 2023). Time-uniform confidence sequences permit repeated looks without a fixed stopping time under their assumptions (Howard *et al.*, 2021). The implemented finite-grid Hoeffding-Ville sequence is intentionally conservative. It is evaluated here as a descriptive monitoring layer over simulator labels, not as a deployment certificate.

MATERIALS AND METHODS

Study design and prospective freeze

The protocol identifier was mafa-integrated-agent-study-v2. Its configuration, benchmark inputs, outcome-affecting source files, prompt, model slots, generation settings, hypotheses, analysis code, and exclusions were committed and tagged before any v2 provider call. The frozen commit was

An operational pilot used eight separate episodes,

7bfa1bc74466b648e7e8d296e2bf5cd6e289fad4; the protocol SHA-256 was 7507ac1a55a1f1102b8589e176a4a9b992e2f92fd02fe29b5a2c512097ea1072; and the frozen source-set SHA-256 was e6c07b4089dd63e5599b1025273f057053e177b43460309dc76a08bb116893e9. This was an internal prospective freeze, not a public preregistration. The main artifact manifest has SHA-256 375d7ea8548d67104ab8f38447cafb77adf569343ec8d65349a67602bfb3443c.

four visible contexts, four model slots, and the first generation seed, for 16 calls. The pilot checked runtime identity, response capture, deterministic replay, and analysis closure. Its outcomes were excluded from the main analysis and were not used to change hypotheses, endpoints, or exclusions. The main run allowed one request attempt, no retry, no replacement, and no output exclusion. Request, telemetry, and parse failures would

have been retained and failed closed. A partially attested logical-model batch would have invalidated the cohort rather than permitting selective resumption.

Synthetic episode matrix

The main benchmark contained 480 episodes across 12 synthetic financial workflow strata. Each stratum contributed ten paired bases, and each pair generated four episode types: stable_safe, unsafe_at_observation, hazardous_interposition, and benign_interposition. Thus there were 120 pair clusters and 40 episodes per stratum. Stable-safe episodes began with valid controls and had no state change before mediation. Unsafe-at-observation episodes exposed an invalid control at the time the model saw the case. Hazardous-interposition episodes began with a valid observation but changed a control and incremented its control-version token before mediation. Benign-interposition episodes changed metadata while leaving the control-version token unchanged.

The model-visible representation omitted the episode identifier, pair identifier, episode type, materiality, expected decision, and outcome label. Workflow text and untrusted context remained visible because they were constant across the four episode types within a paired base. Stable-safe, hazardous-interposition, and benign-interposition variants had byte-identical visible snapshots and shared one proposal context. The unsafe-at-observation variant used a second context because the visible control defect differed. Each pair therefore produced two unique visible contexts, for 240 contexts across 120 pairs.

Episode order was frozen by shuffling rows within each stratum and then emitting strata in rotating round-robin order. Every 12-row block contained one row from each stratum. Consequently, the prespecified audit points at 120, 240, 360, and 480 observations contained equal stratum counts. This property allowed the four static audit schemes to have the same expected 20% label budget at each report point without claiming identical realized audit counts.

Models, proposal generation, and accounting

Four fixed local quantized builds were assigned to logical slots: two Qwen3.5-family builds with 9.7 billion and 27.8 billion parameters, and two Gemma4-family builds with 8.0 billion and 31.3 billion parameters. Each model digest, family, parameter count, quantization (Q4_K_M), installed inventory, and loaded-model attestation was recorded. Annexure A records the exact local tags, digests, prompt, and generation settings. Generation used Ollama 0.32.14, temperature 0.2, a 256-token cap, disabled hidden thinking, sequential execution, and seeds 2026082001 and 2026082002. The accounting followed directly from the paired context design:

- 240 visible contexts × four model slots × two seeds = 1,920 provider calls;

- each model-seed cell mapped those 240 proposals to 480 hidden episode variants;
- 480 proposal-by-episode units $\times$ five mediation arms $\times$ eight model-seed cells = 19,200 transition records.

The final manifest reports 1,920 expected calls, 1,920 completed calls, 1,920 valid calls, and zero request, telemetry, or parse errors. The 16 pilot calls are not included in those totals. Decision agreement between the two seeds, measured on the 240 shared contexts within each model slot, was 91.25%, 92.92%, 96.67%, and 100.00%. These repetitions expose limited generation variation; two seeds do not identify a broader stochastic response distribution.

Mediation arms

Every structurally valid proposal was mapped to its episode variants and replayed through five deterministic arms. The unsafe-at-observation defect was already visible in its separate proposal context. For the valid proposal context, the pipeline then applied no change, a hazardous control change, or a benign metadata change

before mediation. Figure 1 specifies this timing.

- Direct execution committed a structurally valid execute proposal without a policy check.
- Snapshot checking applied deterministic authorization checks to the observation-time snapshot; an approved proposal then committed against live state without revalidation.
- Coarse full-state compare-and-swap required the complete observation state_version to match the live version, then checked and committed against live state. This rejected hazardous state changes but also rejected benign metadata updates.
- Control-state-bound checking required only the observation control-version token to match the live control version, then checked live state and committed atomically.
- Nondeployable authorization reference constructed the authorized requested effect from frozen authorization and live state or abstained. It did not depend on model quality and was included only as a simulator reference.

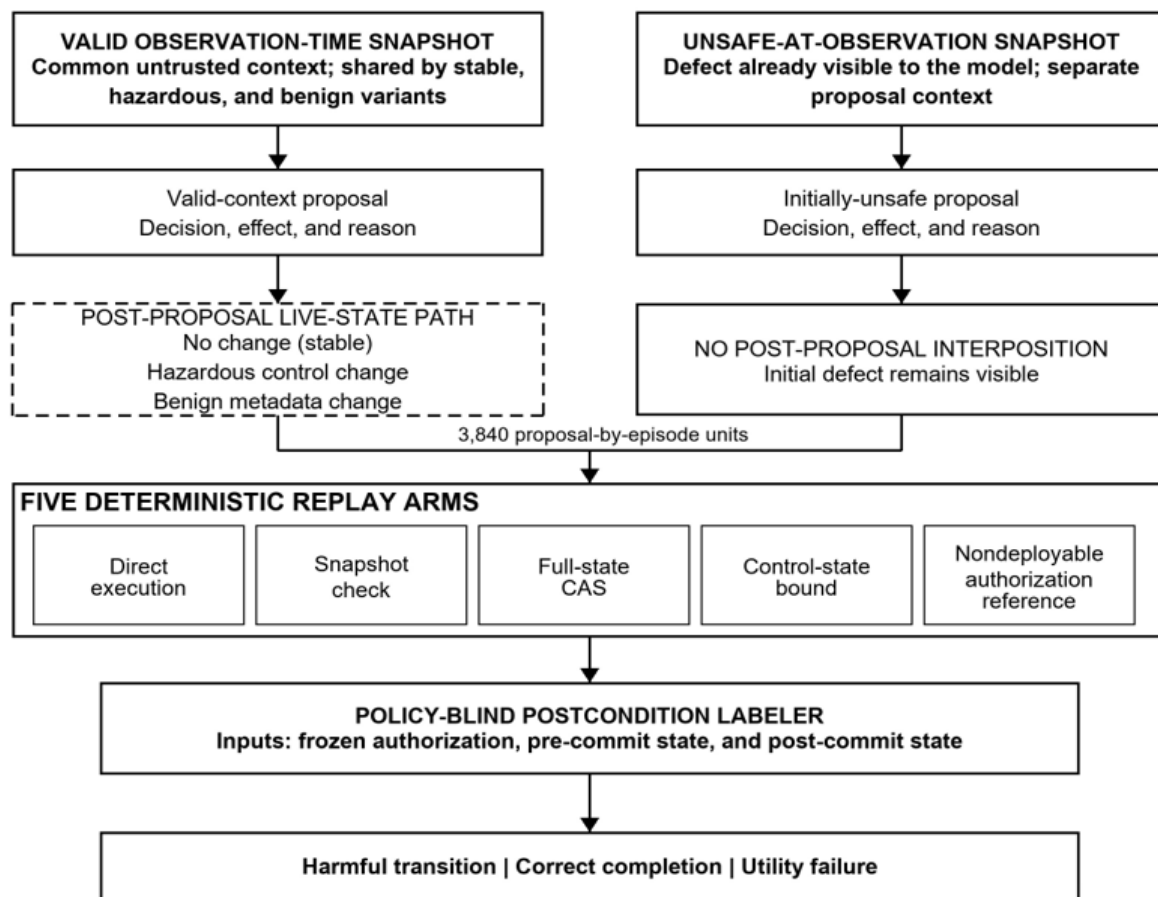


Figure 1: Paired proposal-to-consequence pipeline. Each paired base supplies a valid visible context and a second context in which the unsafe-at-observation defect is already visible. After proposal generation, the valid-context proposal is mapped to no change, a hazardous control change, or a benign metadata change; the initially unsafe context receives no post-proposal interposition. Each proposal-by-episode unit is replayed through five mediators, and a policy-blind postcondition labeler compares frozen authorization, pre-commit state, and post-commit state.

Outcome construction

The primary endpoint was `harmful_state_transition`. Its deterministic labeler accepted only frozen authorization, state immediately before a possible commit, and state after the commit. The labeler could not receive the mediator name, mediator decision, kernel result, expected label, or a model score. Eight rules checked unauthorized execution, wrong beneficiary, amount outside authorization, currency mismatch, duplicate effect, stale-control commit, prohibited disclosure, and ledger non-conservation. Multiple harm types could occur in one transition; the primary endpoint was their union. The authorization kernel compared the proposal with the authorization fields in the state being checked; the postcondition labeler also compared the effect with the frozen originally requested beneficiary and currency. Agreement with mutable authorization-state fields could therefore coexist with a violation of the original request. The frozen request specifies the customer-intended effect for this simulated task. Changes to authorization-state fields are modeled control defects or interpositions, not newly authorized customer instructions. A proposal that follows a changed beneficiary or currency therefore remains incorrect for the original request, even when it agrees with the state fields checked by the kernel. This is an explicit benchmark assumption; a workflow that legitimately replaces the customer's instruction would require a different authorization model and label definition.

Correct completion was a secondary utility endpoint defined only when the live authorization facts made the requested effect eligible. It required the exact authorized effect. Correct abstention described ineligible cases with no effect. Utility failure described an eligible case without correct completion. System nonconformance combined a harmful transition, utility failure, or incorrect abstention. Blocked and malformed proposals did not count as harmful transitions unless state changed harmfully, but they could count as utility failures. This distinction prevents refusal from being credited as both safe and useful.

The postcondition rules were specified by the study authors and tested through deterministic fixtures and mutations. Policy blindness reduces one form of label leakage, but it does not provide human-independent domain validation. The separate authorization reference reimplemented eligibility rather than importing the postcondition labeler; differential fixtures checked agreement. That engineering separation does not turn either implementation into an external ground truth.

Confirmatory estimands and analysis

The target was the balanced synthetic paired-generator distribution conditional on the fixed strata, pair bases, episode variants, models, and seeds. It was not a target over production traffic, financial institutions, future language models, or a statistical population of model families.

H1 was the equal-weight paired risk difference in harmful transitions, `state_bound - direct`, across all four episode variants, models, and seeds. H2 was `state_bound - snapshot_kernel` for harmful transitions on hazardous-interposition episodes. H3 was the correct-completion difference `state_bound - coarse_state_cas` on benign-interposition episodes. H2 was confirmatory only if H1 passed, and H3 only if H1 and H2 passed. For H1 and H2 the criterion was a one-sided 95% generator-sensitivity upper bound below zero; for H3 it was a lower bound above zero. By construction of the nested mediators, every unit-level H1 and H2 contrast is either zero or -1 , while every H3 contrast is zero or $+1$. The confirmatory question is therefore whether a nonzero aggregate magnitude remains separated from zero under the frozen pair-generator resampling plan, not whether the data discover the contrast direction. Gatekeeping orders those magnitude claims; it does not create independent directional evidence.

Each exact finite-suite contrast was accompanied by a stratified generator-population sensitivity analysis. Within each of the 12 strata, the analysis resampled the ten `pair_id` clusters with replacement while retaining all relevant episode variants, models, seeds, and arms for each sampled pair. It used 10,000 replicates and seed 2026082003. Each hypothesis contained 120 pair clusters. The protocol's conservative bounded-outcome calculation gave a one-sided normal-approximation half-width of 0.1502, so the design was suited to moderate or large differences and could be inconclusive for small effects. Model-specific H1 and H2 sign-flip tests were secondary and Holm-adjusted within hypothesis. Model-specific H3 tests were not separately reported because they would duplicate the exact H2 sign-mirror result. The fifth and ninety-fifth bootstrap percentiles form separate one-sided 95% lower and upper bounds; their paired display is not a joint two-sided 95% interval.

The finite-suite estimates are exact for the recorded suite; the bootstrap bounds are sensitivity summaries for variation in pair generation within the fixed strata. Neither is a confidence interval for model populations or real financial events. H2 and H3 also use paired hazardous and benign variants derived from the same visible contexts. Their estimates are algebraically linked when the same proposals execute under one mediator and are blocked under another. They must not be presented as independent replications. The secondary sign-flip calculations were retained as prespecified numerical summaries. The nested arms do not provide exchangeable randomized treatments, so those p-values do not establish independent causal or population evidence.

Selective-audit replay

For each of the $40 \text{ model} \times \text{seed} \times \text{mediator}$ streams, the 480 policy-blind harm labels were ordered by frozen episode index. Four static policies were prespecified for descriptive matched-expected-budget summaries: uniform sampling, materiality-proportional sampling,

a bound-aligned range-optimal allocation, and a full-information variance reference. Each policy had a 5% audit-probability floor and exactly 20% expected labels at the prespecified report points. The full-information reference used the complete realized stratum harm rates from each fixed stream and was nondeployable.

An outcome-history adaptive policy updated stratum priorities from previously audited outcomes. It remained predictable in the simulator, but its realized budget was not constrained to equal the four static policies. It was therefore labeled exploratory and was not used for an equal-budget superiority contrast.

All five schemes used 300 common-random-number replicates per model-seed-arm cell, producing $40 \times 5 \times 300 = 60,000$ audit trajectories. Inverse-probability weighting estimated the full-stream synthetic material-risk target. A finite-grid Hoeffding-Ville confidence sequence used a two-sided delta of 0.05. Report points were 120, 240, 360, and 480. The two prespecified operational summaries were first report point with interval width no greater than 0.10 and first report point with upper bound no greater than 0.05. A trajectory that never crossed by 480 was assigned the censoring time 481. Static outputs were descriptive simulation summaries, not hypothesis tests; H1-H3 were the only confirmatory tests.

Here t counts every stream observation, including unaudited positions. A_i is the audit indicator, π_i its known predictable probability, m_i the synthetic materiality, and Y_i the harm label. The IPW estimate, finite-grid radius, and clipped interval are given below; the two-sided error probability is 0.05. With 17 grid points, the union-bound term is $\log(2 \times 17/0.05) = \log(680)$.

$$\hat{R}_t = \frac{1}{t} \sum_{i=1}^t \frac{A_i m_i Y_i}{\pi_i}$$

$$r_t = \frac{1}{t} \min_{\lambda \in \Lambda} \left\{ \frac{\log(680)}{\lambda} + \frac{\lambda}{8} \sum_{i=1}^t \left(\frac{m_i}{\pi_i} \right)^2 \right\}$$

$$\Lambda = \{2^k : k = -12, \dots, 4\}$$

$$C_t = [\text{clip}_{[0,1]}(\hat{R}_t - r_t), \text{clip}_{[0,1]}(\hat{R}_t + r_t)]$$

$$\text{clip}_{[0,1]}(x) = \max\{0, \min\{1, x\}\}$$

The bounded Hoeffding construction gives an explicit finite-grid calculation but pays for worst-case ranges. It was specified in the frozen protocol; its conservatism is retained in the reported analysis. Annexure A gives the allocation rules and materiality sensitivity settings.

Integrity controls and data availability

Before each logical-model batch, the runner authenticated the installed model inventory and decoding settings. After each batch, it rechecked the inventory and verified that the

intended model digest was loaded. Raw response bytes and complete provider envelopes were flushed and file-synchronized before a durable completion event. Journal records bound the commit, tag, protocol hash, source-set hash, prompt hash, model digest, seed, context hash, evidence-file hashes, and runtime attestation. Frozen source and input bytes were rechecked before replay and after analysis. The complete artifact root became nonresumable after finalization.

The authenticated manifest records Python 3.11.14 for the completed run. The repository root pins the same interpreter in `.python-version`; reproduction and validation should stop if `python --version` does not report Python 3.11.14. The executable benchmark, analysis, and validation paths use the Python standard library and introduce no external Python package dependency.

The authenticated aggregate artifacts needed to reproduce the reported numerical claims are the manifest, hypothesis contrasts, arm summary, model-specific contrasts, transition records, audit summary, and audit replicates. The claim-to-artifact map distributed with this source records paths, filters, denominators, and transformations. Raw model text and provider envelopes are not required to verify the aggregate paper claims and should undergo separate licensing, privacy, and content screening before any public release. Hashes establish byte integrity within this workflow; they do not establish author identity, hardware trust, or independent replication. The accompanying reproduction guide provides an offline replay route for Python 3.11.14 using the recorded proposals, synthetic episodes, and analysis code. It separates checks that work from an extracted archive from provenance checks that require Git metadata. The Word supplements provide evidence in six parts (04A–04F), source with its original license (05), and an extraction guide (07). The source file is identifying and requires editorial handling before reviewer distribution. Annexure A supplies the essential methods in the article. Administrative-only protocol text is omitted in a documented review derivative; frozen scientific inputs and results remain unchanged.

RESULTS AND DISCUSSION

Run closure and descriptive arm outcomes

The main study completed all 1,920 planned calls and all 19,200 transition replays. No request, telemetry, or parse failure occurred. This complete-case result means that the prespecified failure-retention policy was not exercised; it provides no empirical evidence about behavior under transport failure, malformed provider envelopes, partial writes, or batch invalidation.

Table 1 reports aggregate consequences across 3,840 proposal-by-episode units per arm. State-bound checking and coarse full-state checking each produced 24 harmful transitions (0.63%). Snapshot checking produced 808 (21.04%), direct execution 1,126 (29.32%), and the nondeployable reference zero. Among 1,920 live-eligible units per arm, state-bound, snapshot, and direct execution

Table 1: Aggregate outcomes by mediation arm.

Mediation arm	Harmful transitions, n/3,840	Harm rate	Correct eligible completions, n/1,920	Completion rate	System nonconformance rate
Direct execution	1,126/3,840	29.32%	1,568/1,920	81.67%	38.44%
Snapshot checking	808/3,840	21.04%	1,568/1,920	81.67%	30.21%
Coarse full-state compare-and-swap	24/3,840	0.63%	784/1,920	40.83%	30.21%
Control-state-bound checking	24/3,840	0.63%	1,568/1,920	81.67%	9.79%
Nondeployable authorization reference	0/3,840	0.00%	1,920/1,920	100.00%	0.00%

each completed 1,568 (81.67%). Coarse full-state checking completed 784 (40.83%) because it rejected every benign metadata interposition. The authorization reference completed all 1,920 eligible units by construction.

The episode-type breakdown in Table 2 explains the aggregate rates. Snapshot checking's 784 hazardous-interposition harms occurred because approval against the old snapshot did not protect the live commit. Direct execution had 785 such harms and 339 harms on cases already unsafe at observation, plus one harm in each of

the stable and benign variants. State-bound and coarse checking eliminated hazardous-interposition harms but retained 24 harms in unsafe-at-observation episodes. All 24 residual state-bound harms involved a mismatch with the frozen requested effect: 16 involved currency mismatch and eight involved a wrong beneficiary. In these cases, the proposal matched the visible authorization-state fields checked by the kernel but differed from the frozen request used by the postcondition labeler. Checking state freshness and conformance to those fields

Table 2: Harmful transitions by episode type and mediation arm (960 proposal-by-episode units per arm-by-type cell).

Mediation arm	Stable safe	Unsafe at observation	Hazardous interposition	Benign interposition
Direct execution	1	339	785	1
Snapshot checking	0	24	784	0
Coarse full-state compare-and-swap	0	24	0	0
Control-state-bound checking	0	24	0	0
Nondeployable authorization reference	0	0	0	0

did not establish conformance to the original request. This residual prevents the result from being summarized as complete prevention.

The system-nonconformance endpoint adds utility failure and incorrect abstention to harm. State-bound checking had the lowest nonreference rate, 9.79%, because it accepted benign state changes while checking the current control state. Coarse checking and snapshot checking both reached 30.21% through different mechanisms: coarse checking lost eligible benign completions, whereas snapshot checking allowed hazardous commits. Equal aggregate nonconformance therefore does not imply equal error composition. A deployment choice would

require domain-specific costs for these errors, which the study did not elicit.

Proposal-level behavior

Table 3 separates the empirical proposal decisions from their mediator-dependent consequences. Each build contributes 240 valid-context and 240 unsafe-at-observation proposals across its two fixed seeds. The stable, hazardous, and benign variants share the valid-context proposal and are not three independent model decisions. Across builds, 785 of 960 valid-context proposals requested execution (81.77%); 784 were correct and passed the kernel (81.67%).

Table 3: Recorded proposal decisions and residual harms by fixed model build. Execute/other gives the two decision counts within each 240-proposal context group. Approved counts are correct valid-context executions. Unsafe blocked counts include only execute proposals rejected by the kernel; residual harms are among unsafe-at-observation state-bound transitions.

Recorded build	Valid execute / other	Valid approved	Unsafe execute / other	Unsafe blocked	Residual harm
qwen3.5:9b	177 / 63	176	111 / 129	91	20
qwen3.5:27b	200 / 40	200	62 / 178	62	0

gemma4:e4b	186 / 54	186	104 / 136	100	4
gemma4:31b	222 / 18	222	62 / 178	62	0
Total	785 / 175	784	339 / 621	315	24

The extra valid-context execute was a qwen3.5:9b card-refund proposal with the correct amount and currency but an incorrect beneficiary identifier. Direct execution therefore produced one harm in each stable and benign variant and the 785th hazardous-interposition harm; the checking arms rejected it. Of the 24 residual state-bound harms, qwen3.5:9b contributed 16 currency and four beneficiary mismatches, and gemma4:e4b contributed four beneficiary mismatches. These are results for recorded builds and two seeds, not estimates of model-family performance.

The H1 harm reduction can consequently be decomposed into 315 blocked unsafe-context executions, 785 prevented hazardous commits, and the two stable/benign errors: $(315 + 785 + 1 + 1)/3,840 = 1,102/3,840$. The aggregate magnitude depends on the chosen episode mixture and the observed proposal behavior; the equal four-way design gives each episode type one quarter of the total weight. The nested mediator definitions constrain the direction.

Confirmatory contrasts

All three gatekeeping criteria were met in the frozen suite (Table 4). This does not mean that an unconstrained direction was discovered: the mediator definitions restrict H1 and H2 to nonpositive unit contrasts and H3 to nonnegative unit contrasts. The result is that the prespecified bounds place the generator-resampled aggregate magnitude away from zero. For H1, state-bound checking produced 24 harms among 3,840 units compared with 1,126 under direct execution. The paired risk difference was -0.2870 , and the separate one-sided 95% generator-sensitivity bounds were -0.2901 to -0.2836 . Because the upper bound was below zero, H1 passed and opened H2.

For H2, state-bound checking produced no harmful transition among 960 hazardous-interposition units, compared with 784 under snapshot checking. The paired difference was -0.8167 , with bounds -0.8229 to -0.8104 . H2 passed and opened H3. For H3, state-bound checking correctly completed 784 of 960 benign-

Table 4: Prespecified gatekept paired contrasts.

Hypothesis contrast	Arm counts	Paired difference	One-sided 95% generator-sensitivity bounds	Pair clusters	Criterion met
H1: state-bound – direct harmful transition, all units	24/3,840 vs. 1,126/3,840	-0.2870	$[-0.2901, -0.2836]$	120	Yes
H2: state-bound – snapshot harmful transition, hazardous interposition	0/960 vs. 784/960	-0.8167	$[-0.8229, -0.8104]$	120	Yes
H3: state-bound – coarse correct completion, benign interposition	784/960 vs. 0/960	$+0.8167$	$[+0.8104, +0.8229]$	120	Yes

interposition units, compared with zero under coarse full-state checking. The paired difference was $+0.8167$, with bounds $+0.8104$ to $+0.8229$.

In the prospectively frozen controlled simulator, state-bound mediation reduced the tested-suite harmful-transition rate by 28.70 percentage points relative to direct execution. On hazardous interpositions it reduced harm by 81.67 points relative to snapshot checking, while on the paired benign interpositions it increased correct completion by 81.67 points relative to coarse full-state CAS. The latter two contrasts are deterministic sign mirrors over shared proposal contexts and should not be interpreted as independent replications.

The sign-mirror qualification matters. The hazardous and benign variants share proposal contexts. In the relevant cells, the same execute proposal that snapshot checking carries into a hazardous live state is stopped by state-

bound checking; the corresponding benign variant is accepted by state-bound checking and stopped by full-state version checking. The identical absolute H2 and H3 estimates therefore arise from the paired mechanism and proposal set. Reporting them as two independent confirmations would overstate the evidence.

Secondary model-specific H1 estimates ranged from -0.2729 to -0.2979 ; H2 estimates ranged from -0.7333 to -0.9250 . Each model-specific sign-flip calculation reached the add-one Monte Carlo resolution floor, $1/10,001$ (approximately 0.0001), with a Holm-adjusted value of approximately 0.0004 within hypothesis. The observed per-build estimates show that no one logical slot solely produced the aggregate direction. Because the nested arms are not exchangeable randomized treatments, the small sign-flip p-values do not provide an independent confirmation of that direction or support a causal or

population claim. These descriptive comparisons do not rank models. Model identity is inseparable from fixed build, quantization, and sequential run order, and only two seeds were used.

Selective-audit behavior

Table 5 summarizes the prespecified endpoint at 480 observations, averaged descriptively over the 40 fixed model-seed-arm cells. The four static schemes each had an expected audit rate of exactly 0.20. Their pooled mean realized rates ranged from 0.2000 to 0.2003. The adaptive scheme's pooled mean realized rate was 0.2003, but its expected rate varied with outcome history and it is not an equal-budget comparator.

Table 5. Descriptive selective-audit summaries at 480 observations. Mean absolute error is the absolute difference between the IPW estimate and the complete-label material-weighted harm proportion in the same 480-event stream, averaged over 300 replicates and 40 model-seed-mediator cells. It measures audit estimation error, not model prediction accuracy. Width is the clipped confidence-interval width.

Across all 60,000 trajectories, no scheme crossed the prespecified interval-width threshold of 0.10 or upper-bound threshold of 0.05 at 120, 240, 360, or 480 observations. Every time-to-threshold value was therefore censored at 481. Finite-grid point and prefix coverage were 100% in every displayed cell and report

Table 5: Descriptive selective-audit summaries at 480 observations. Mean absolute error is the absolute difference between the IPW estimate and the complete-label material-weighted harm proportion in the same 480-event stream, averaged over 300 replicates and 40 model-seed-mediator cells. It measures audit estimation error, not model prediction accuracy. Width is the clipped confidence-interval width.

Audit scheme	Status	Cells	Mean realized audit rate	Mean expected audit rate	Mean absolute error	Mean finite-grid width
Uniform	Prespecified static	40	0.2000	0.2000	0.0093	0.3228
Materiality proportional	Prespecified static	40	0.2003	0.2000	0.0083	0.2680
Range optimal	Prespecified static	40	0.2003	0.2000	0.0082	0.2639
Full-information variance reference	Prespecified static, nondeployable	40	0.2001	0.2000	0.0064	0.5179
Adaptive	Exploratory, realized rate only	40	0.2003	0.2001	0.0082	0.2762

point. This perfect empirical coverage should not be read as calibration superiority: the intervals were conservative enough that none produced the intended operational certificate within the horizon. The absence of crossing is a central result, not an implementation inconvenience to tune away after observing outcomes.

Inspection of the implemented bound explains why non-crossing at this horizon is unsurprising. At 480 observations, uniform sampling has 96 expected audits but a deterministic interval radius of 0.2732; its pooled clipped width is 0.3228. Materiality and range-optimal sampling have radii 0.2182 and 0.2141, respectively. Even a zero-harm uniform stream therefore has an upper bound about 0.2732, above the required 0.05, and width above 0.10. This is a post-hoc analytical explanation of the prespecified procedure, not an a priori horizon calculation or a new experiment.

Betting-based financial-audit methods offer a motivated comparison for future work (Shekhar *et al.*, 2023). Their weighted sampling without replacement targets a finite population, whereas this replay uses predictable Bernoulli selection and sequential-prefix material risk. An adapted estimator and coverage argument would be needed before

comparing stopping behavior. No such comparison was run, and the frozen Hoeffding analysis has not been replaced retrospectively.

Pooled mean width was smallest for range-optimal allocation and largest for the full-information variance reference, but these are descriptive averages across heterogeneous consequence streams, not controlled superiority estimates. The variance reference optimizes a different synthetic objective and can assign probabilities that interact poorly with the finite-grid worst-case range. The protocol specified no audit hypothesis, superiority margin, or multiplicity correction. Consequently, the table supports implementation diagnosis and future design, not a headline claim that one audit policy is best.

Interpretation and limitations

The paired results isolate one simulator mechanism: when control state changes after observation, checking the live control-version token and current state can prevent stale commits without rejecting metadata-only changes. Direct execution and snapshot checking lack that interposition protection; full-state compare-and-swap protects it at a utility cost. This is evidence about how established

patterns behaved in a fixed state machine, not evidence that the patterns are new or that one implementation is ready for use.

The synthetic design is the largest boundary. The 12 workflow strata, authorization facts, interpositions, reference probabilities, and state transitions were constructed. They do not estimate the prevalence, cost, or severity of events in a bank, payments company, insurer, investment firm, or public agency. No customer data, production logs, settlement rails, sanctions screening service, or institution policy was used. External validity requires domain review, governed data, and an evaluation inside a controlled real workflow.

Materiality is also synthetic and normatively incomplete. A scalar weight cannot by itself represent finality, privacy, discrimination, sanctions exposure, customer harm, liquidity, or operational disruption. The analysis includes declared sensitivity exponents, but changing an exponent does not validate the underlying weights. No expert panel reviewed them, and discovered synthetic material harm should not be translated into dollars or regulatory significance.

The outcome labeler is policy-blind with respect to mediator identity, but it is not independent ground truth. The authors defined the authorization facts, state machine, and eight postcondition rules. Differential and mutation tests support implementation consistency, not domain correctness. The nondeployable authorization reference also succeeds by construction from simulator authorization and live state. Its zero harm and complete eligible utility are a ceiling check, not a realistic operational benchmark.

Model evidence is limited to four fixed local quantized builds, one prompt, one decoding configuration, two seeds, and sequential execution on one machine. The suite does not support a population claim about Qwen, Gemma, open models, proprietary systems, or future checkpoints. Decision agreement of 91.25%–100.00% across two seeds shows that some proposal variation remained, but two draws are too few to characterize generation variance. Model-specific tests are secondary and confounded with build and run order.

The confirmatory analysis has structural dependencies. H1-H3 share pair clusters, models, seeds, and proposals; gatekeeping controls the declared sequence but does not create independent evidence. More fundamentally, the nested mediator definitions constrain H1 and H2 unit contrasts to zero or -1 and H3 unit contrasts to zero or $+1$. The tests quantify whether the resulting aggregate magnitudes remain separated from zero under generator resampling; they are not empirical tests of a freely varying sign. H2 and H3 are also exact sign mirrors over shared contexts. The bootstrap resamples generated pairs within fixed strata and therefore describes sensitivity to that generator, not uncertainty over institutions, tasks, or model families. With 120 clusters and a conservative planning half-width of 0.1502, the study was not designed to resolve small effects.

State-bound checking retained 24 harms, all in unsafe-at-observation episodes. This is expected from the intervention's scope: freshness can detect a changed control token but cannot repair an authorization defect that was already present. The method also assumes that the selected control token captures every relevant execution-time condition and that the check and commit are atomic. Real systems can have distributed state, partial failures, cross-service races, inconsistent clocks, and bypass paths not represented here.

The run completed with zero provider failures. That supports clean accounting but leaves fail-closed behavior under transport errors, length termination, missing telemetry, malformed envelopes, or runtime changes unobserved as study outcomes. Tests exercise those branches, but fixture tests are not provider observations. Similarly, hashes and attestations protect internal provenance; they do not provide cryptographic user authorization, trusted hardware, or external replication.

The audit replay assumes known positive predictable sampling probabilities, prompt access to the harm label when audited, and no outcome-dependent missingness outside the policy. Complaint-driven review, delayed settlement, auditor discretion after seeing a result, label correction, or unavailable evidence can violate those assumptions. All four static schemes were compared only descriptively at matched expected budgets, not identical realized budgets. The adaptive scheme was exploratory and not equal-budget. Three hundred replay replicates per cell provide finite simulator resolution but do not establish operational coverage. Most importantly, no trajectory reached either stopping threshold by 480 observations.

Finally, the endpoint set is incomplete. It includes eight state-transition harms and eligible completion, but it does not measure fairness, explanation quality, legal sufficiency, cybersecurity beyond the specified state machine, macroeconomic impact, latency, implementation cost, or human factors. The study is not legal, accounting, investment, security, or compliance evidence. Any deployment path would require independent scenario review, threat modeling, control verification, shadow operation, human-approved limits, incident procedures, and fresh evaluation under the intended environment.

CONCLUSION

This controlled paired study evaluated how five execution arrangements translated the same local-model proposals into consequences under stable, initially unsafe, hazardously changed, and benignly changed synthetic financial states. The internally frozen main run completed 1,920 calls and 19,200 transition records without exclusions or provider errors. In the tested suite, control-state-bound checking reduced harmful transitions relative to direct execution and snapshot checking while preserving benign-interposition completion relative to coarse full-state checking. The mediator definitions structurally constrained those directions, so the prespecified bounds

quantify nonzero magnitude under generator resampling rather than discovering a sign. State-bound checking did not eliminate 24 harms already present at observation, and its two interposition contrasts were algebraic sign mirrors rather than independent confirmations.

The selective-audit replay did not produce an operational stopping certificate. Although realized audit rates tracked the intended 20% level and empirical coverage was 100% in displayed cells, every one of 60,000 trajectories remained above both prespecified stopping thresholds through 480 observations and was censored at 481. This result favors reporting conservative monitoring limits honestly rather than converting coverage alone into a readiness claim.

The appropriate conclusion is bounded: binding an action to current control state changed measured post-state consequences in this fixed simulator and preserved utility against metadata-only interpositions, while the selected audit confidence procedure remained too wide for the prespecified stopping rules. The next evidentiary step is not another claim about mechanism novelty. It is independent financial-domain review, a newly frozen external replication with broader models and failure modes, and validation in a governed shadow workflow before any operational inference.

Ethics statement

This computational study used synthetic financial-workflow records and involved no human participants, animals, identifiable customer data, or production financial records. Informed consent was not applicable. No institutional ethics approval or exemption is claimed.

REFERENCES

- Chen, H.-H. (2026). *Insuring every action: An authority frontier framework for runtime actuarial control of autonomous AI agents* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2605.25632>
- Chen, Z., Chen, J., Chen, J., & Sra, M. (2025). *Standard benchmarks fail—Auditing LLM agents in finance must prioritize risk* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2502.15865>
- Chen, Z., Chen, J., Chen, J., & Sra, M. (2026). From tasks to teams: A risk-first evaluation framework for multi-agent LLM systems in finance. In M. Liakata, V. P. Moreira, J. Zhang, & D. Jurgens (Eds.), *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 38819–38857). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.findings-acl.1934>
- DeBenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in neural information processing systems* (Vol. 37, pp. 82895–82920). Curran Associates. <https://doi.org/10.52202/079017-2636>
- Feng, Y., Lin, R., Wen, M., He, Q., Guo, Y., Ding, Y., Wu, Y., Chen, J., Xu, Z., Du, X., Ma, J., Chen, Z., Ma, X., Chen, Y., & Deng, X. (2026). *Safety testing LLM agents at scale: From risk discovery to evidence-grounded verification* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2607.01793>
- Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685. <https://doi.org/10.1080/01621459.1952.10483446>
- Hou, Y., Jiang, Y., Xie, Y., Yang, J., Zhang, L., Huang, H., Chen, G., & Chen, Y. (2026). FinSafetyBench: Evaluating LLM safety in real-world financial scenarios. In M. Liakata, V. P. Moreira, J. Zhang, & D. Jurgens (Eds.), *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 14181–14208). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.findings-acl.694>
- Howard, S. R., Ramdas, A., McAuliffe, J., & Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 1055–1080. <https://doi.org/10.1214/20-AOS1991>
- Huang, D., Chua, J. K., & Wang, Z. (2026). *Beyond task success: Measuring workflow fidelity in LLM-based agentic payment systems* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2605.06457>
- Li, J., & Zhu, J. (2026). *Auditing self-evolution in financial agents: Capability gains, security drift, and execution-interface mismatch* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.17684>
- Lu, J., Wang, K., Wang, Y., Tang, Q., Zeng, H., Chen, X., Pi, J., Deng, S., Chen, L., Fu, Y., Yang, K., & Sun, X. (2026). *FinToolBench: Evaluating LLM agents for real-world financial tool use* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2603.08262>
- Lu, Z., Zuo, Y., Xu, H., Chen, W., He, X., Guo, J., & Jin, S. (2026). *WirelessOpsAgent: A benchmark and agent design for action assurance in wireless networks* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.08277>
- Luo, M., Chen, C., Cao, A., Huang, Z., & Dai, W. (2026). *STAGE: Stateful translation to agentic graph execution with policy-scoped context and deterministic control* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.22538>
- Luo, Y., Jiang, Y., Xie, Q., Lan, L., Cong, L. W., Rao, A., & Song, Y. (2026). *ReguSim: Evaluating LLM agent rule grounding in financial compliance* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.19974>
- Mao, Q., Wang, J., Liu, Y., Zhu, L., Ma, C., & Yan, J. (2026). *SoK: Security of autonomous LLM agents in agentic commerce* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2604.15367>
- Meng, F., Du, L., Wu, Z., Chen, G., Liu, X., Liao, J., Jiang, C., Wan, Z., Gu, J., Zhou, P., Huang, R., Zhao, Z., Ding, S., Yu, A., Peng, B., Xia, B., Sun, H., Liang, H., Xie, J., ... Shieh, M. Q. (2026). *ClawMark: A living-world benchmark for multi-turn, multi-day, multimodal coworker agents* [Preprint]. arXiv. <https://doi.org/10.48550/>

- arXiv.2604.23781
- Mohammed, Y., & Lund, B. (2026). Cultural-historical activity theory and AI: Innovating and optimizing financial data retrieval. *American Journal of Financial Technology and Innovation*, 4(1), 75–89. <https://doi.org/10.54536/ajfti.v4i1.4187>
- Mou, Y., Xue, Z., Li, L., Liu, P., Zhang, S., Ye, W., & Shao, J. (2026). ToolSafe: Enhancing tool invocation safety of LLM-based agents via proactive step-level guardrail and feedback. In M. Liakata, V. P. Moreira, J. Zhang, & D. Jurgens (Eds.), *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 37125–37153). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.findings-acl.1850>
- Pai, D. M., & Xian, L. (2026). *FraudBench: Stress-testing policy-grounded banking agents against adaptive fraud* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.18136>
- Peng, Y., & Wu, X. (2026). *Stateful governance for concurrent agentic systems* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.02764>
- Plaat, A., van Duijn, M., van Stein, N., Preuss, M., van der Putten, P., & Batenburg, K. J. (2025). Agentic large language models, a survey. *Journal of Artificial Intelligence Research*, 84, Article 29. <https://doi.org/10.1613/jair.1.18675>
- Raidah, F. Z., Jobair, M., Halimuzzaman, M., Sharma, J., & Ahmed, S. S. (2026). Artificial intelligence and the transformation of internal audit functions. *American Journal of Financial Technology and Innovation*, 4(1), 178–184. <https://doi.org/10.54536/ajfti.v4i1.8001>
- Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C. J., & Hashimoto, T. (2024). Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/7274ed909a312d4d869cc328ad1c5f04-Abstract-Conference.html
- Santos-Gruciro, I. (2026). *Temporary authority, permanent effects: Commit-time authorization for LLM agents* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2607.10487>
- Shekhar, S., Xu, Z., Lipton, Z., Liang, P., & Ramdas, A. (2023). Risk-limiting financial audits via weighted sampling without replacement. In R. J. Evans & I. Shpitser (Eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence* (Proceedings of Machine Learning Research, Vol. 216, pp. 1932–1941). PMLR. <https://proceedings.mlr.press/v216/shekhar23a.html>
- Shi, T., Mo, Y., Liu, Y., Hao, Z., Wang, Y., Hu, W., Yu, N., Zhou, M., & Yu, J. (2026). *Organizational control layer: Governance infrastructure at the execution boundary of LLM agent systems* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2606.04306>
- Tang, R., Liu, Q., Zhang, Y., Yang, Y., Chen, X., & Dong, C. (2026). *Context is not authority: Structured runtime governance for financial market agents* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.09025>
- Tong, Y., Dai, L., & Guo, S. (2026). *AID-Guard: Stateful authorization for delegated agent effects* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.21159>
- Wang, H., Poskitt, C. M., & Sun, J. (2025). *AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2503.18666>
- Xia, H., Wang, H., Liu, Z., Yu, Q., Guo, Y., & Wang, H. (2025). SafeToolBench: Pioneering a prospective benchmark to evaluating tool utilization safety in LLMs. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 17643–17660). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-emnlp.958>
- Xu, J., Fan, L., Wang, Z., Li, X., & Liu, H. (2026). *Beyond single-use tokens: Durable authorization state for replay-resistant LLM agent actions* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.01710>
- Yang, Z., Li, R., Qiang, Q., Wang, J., Lou, F., Li, M., Cheng, D., Xu, R., Lian, H., Zhang, S., Liang, X., Huang, X., Wei, Z., Liu, Z., Guo, X., Wang, H., Chen, R., & Zhang, L. (2026). *FinVault: Benchmarking financial agent safety in execution-grounded environments* [Preprint, version 1]. arXiv. <https://arxiv.org/abs/2601.07853v1>
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=WE_vluYUL-X