BINDING USER APPROVAL TO INDIVIDUAL ACTIONS IN TOOL-USING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tool-using agents often ask users to approve an action before executing it. However, when an identical action is proposed again, matching the tool, target, and arguments does not establish which proposal was approved. We introduce EPOCHSELECT, a runtime contract and reference adapter that bind approval to a specific action instance: the tool's preview path must create the instance before approval, approval can only select exactly one existing live instance, and execution rechecks its live state and final arguments. Under explicit synchronous-execution and trust assumptions, the contract prevents approval and execution evidence from being combined across different instances. In 480 matched in-process restoration tests, full instance binding accepts all 120 exact restorations and rejects all 360 nonexact rebindings, avoiding the valid restorations blocked by the object-identity baseline and the invalid substitutions admitted by the value-matching baseline. Language-level evaluation reveals a complementary limitation: on 128 proxy turns from benchmark training data with real tool handlers and preselected proposals, a repaired approval rule rejects all 64 negative cases but accepts only 25 of 64 positive cases (39.1%), each executed exactly once. These results support instance-level approval binding in the tested setting, while leaving natural-language target resolution and end-to-end task effectiveness unestablished.

1 INTRODUCTION

Preview/confirm interfaces let users inspect a tool action before authorizing execution. Matching its displayed value is insufficient when the same action can recur. Suppose a travel agent previews cancellation of reservation r, that obligation ends, and a value-identical cancellation is proposed again. Approval must identify one occurrence, not the shared cancellation value. If authorization code constructs the occurrence after reading approval, it can manufacture a consistent object that did not exist when the user acted. Its digest proves self-consistency, not prior existence. This problem concerns stateful tool interfaces such as those exercised by agent benchmarks (Yao et al., 2023; Ruan et al., 2024; Debenedetti et al., 2024; Yao et al., 2024).

We call this failure *authority-to-execution instantiation* and address the ownership of proposal creation:

> **A native handler creates a proposal occurrence; authority may only select an occurrence that the mediation surface has already opened.**

Our reference adapter, EPOCHSELECT, records a verified native proposal before approval. Later authority selects exactly one live occurrence; it cannot create one. Before the effect, a guarded dispatch replays the ledger and revalidates the occurrence's live sources and final typed arguments.

Claim and scope. For a synchronous adapter satisfying Contract C (Section 2), every successful write is tied to the unique handler-opened occurrence selected by authority. Implementation evidence covers two frozen entry roots and finite generated designs, not arbitrary runtimes or natural-language target resolution. Occurrence integrity and language coverage are distinct: a correct binding is unusable if the consent gate rejects ordinary approvals, while a fluent classifier cannot establish that the approved occurrence existed beforehand.

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

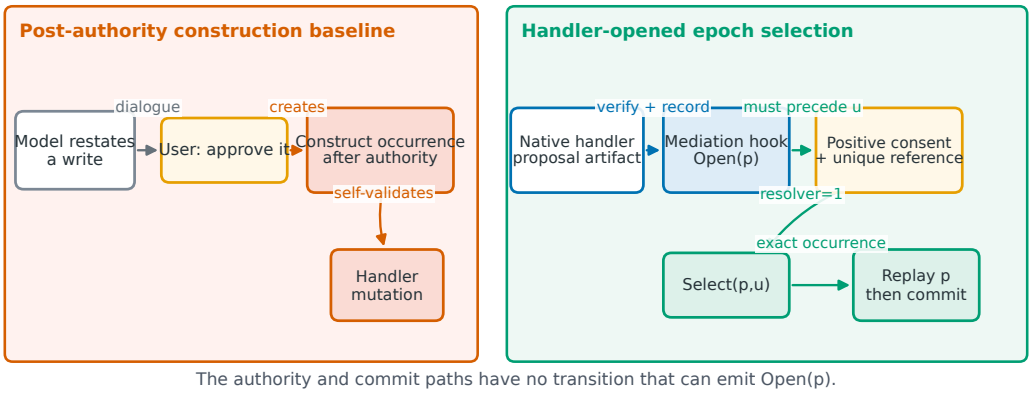

Figure 1: **Transaction ownership.** An authority-time guard can construct the object it later validates (left). EPOCHSELECT lets only the native proposal path create and surface a complete occurrence; authority selects that prior object (right).

Prior systems govern tool-call policy (Uchibeke, 2026), trusted rendering (Weng, 2026), action envelopes and fingerprints (Wang, 2026b;a), authorization–execution histories (Rhodes & Kang, 2026), or pre-effect decisions (Mazzocchetti, 2026). Such systems can carry occurrence identifiers. Our contribution is the adapter condition that makes an identifier evidence of a prior native occurrence: only the proposal path may create it, and selection and dispatch must return to its live sources. Proposition 1 identifies the required discriminating information; it does not rule out existing action envelopes adding that information. Appendix I compares these abstractions.

Our contributions are:

1. an occurrence-owned transition contract and conditional invariants for prior opening, live uniqueness, mediated dispatch, and splice rejection;
2. frozen conformance studies spanning a reference adapter, named integration shapes, and 480 matched in-process restore workloads, where full binding accepts 120 exact restores and rejects 360 nonexact cases; and
3. language-interface diagnostics that distinguish mechanism conformance from usability: repaired gating accepts only 39.1% of positive proxies despite target preselection, and a full-history census supplies no executable comparison under its fixed adapter and grammar.

The last result bounds the tested interface, not the target supply of the source benchmarks; neither diagnostic establishes end-to-end task efficacy.

2 DELAYED WRITES ARE OCCURRENCE-SELECTION PROBLEMS

Native interface. Let a native write tool expose a non-mutating preview handler and a mutating commit handler. A preview produces a typed proposal artifact q containing the intended tool call, target, arguments, and native admission metadata. A mediation surface renders q to the user. The surface event is observable and deterministic; model prose alone is not a proposal witness.

Threat model. The model may repeat or paraphrase a proposal, issue an invalid request, reference an old proposal, or continue after dependencies, target authority, or obligation lifecycle have changed. The user may negate, question, conditionally promise, quote, revoke, or ambiguously express authority. The adversary controls these untrusted inputs and may replay, reorder, delete, or coherently rewrite every serialized certificate, ledger record, proposal snapshot, transcript field, and unkeyed hash. The adversary may not modify a trusted verifier, a live native runtime object, or the sole dispatch entry.

The trusted computing base is explicit: native typed admission and dependency verifiers, live native runtime objects, the proposal-surface hook, the ledger replayer, and an assumed guarded dispatch entry. A digest establishes identity only after these trusted verifiers re-enter and validate the serialized object. A misleading trusted rendering, a cryptographic collision, or an effectful path that bypasses the dispatch entry is outside the model and is not silently converted into an empirical security claim.

Contract C (the refinement boundary). The specification invariants assume five obligations: (C1) only the verified proposal-surface hook can append an opening; (C2) every authority and commit decision consumes a unique live occurrence whose complete source bundle has been revalidated; (C3) every effectful native write is invoked only through a single dispatch hook that rechecks the final typed arguments and occurrence identity; (C4) no external yield or concurrent mutator occurs between final replay and the handler call; and (C5) the native handler and the live runtime objects implement the typed schemas used by the verifiers. These invariants characterize compliant traces; they do not establish that an arbitrary program satisfies C1–C5. The primary contract binds a 375-entry source closure. A post-run audit re-hashes nine C3-relevant files and discharges C3 only for the two frozen primary-study entry roots. For Study E, a later explicit-taint audit freshly parses all 79 current source modules and 11 declared runtime roots, classifies 689 handler-flow events, and finds the normalized core signature unchanged after the authority hook. This is syntactic coverage under a frozen grammar, not a whole-program proof: dynamic dispatch, C4–C5, and undeclared effectful paths remain premises.

Execution contract. Contract C models the scored adapter as single-process and synchronous and requires all effectful commit branches to route the final native call through one $\mathsf{Dispatch}(p)$ hook. The hook rechecks the selected cardinality, tool/target identity, opening and source snapshot, and the typed hash of the post-transform arguments immediately before invoking the handler; a failed check returns a governance-blocked result without calling the handler. C4 assumes that no external yield or concurrent mutator occurs between replay and dispatch. This is the atomic validation-use boundary required by the invariants, rather than an attempt to solve generic time-of-check/time-of-use races (Dean & Hu, 2004). An asynchronous adapter must provide an equivalent lock or versioned atomic dispatch primitive and must re-run the same live-source checks after a failed version check; otherwise the invariants do not apply.

Throughout, *single-dispatch* is a contract assumption about writes in the claimed adapter scope: every such effectful write is routed through one synchronous final path in this guarded hook. It does not mean one total low-level handler invocation per episode: a native preview/confirm adapter may make a non-mutating preview call before the later effectful confirmation call.

Why values are insufficient. Let two obligations o_1 and o_2 request the same tool, target, and business arguments. They are nevertheless distinct occurrences when they have different obligation identifiers, generations, or issued-event hashes. A value-only approval key collapses o_1 and o_2 and permits stale authority, dependency evidence, or retirement evidence to migrate between them.

Proposition 1 (Occurrence-oblivious indistinguishability)**.** *Let an authority selector observe only the user act and $\pi(p) = (\mathit{tool}, \mathit{target}, \mathit{typed\ arguments})$. For two distinct occurrences $p_1 \neq p_2$ with $\pi(p_1) = \pi(p_2)$ and the same authority act, every deterministic selector over that observation gives the same decision. It therefore cannot accept the intended live occurrence while rejecting a stale binding to the other occurrence.*

Proof. The selector's complete input is identical in the two executions, so determinism gives identical output. Distinguishing the required decisions needs at least one occurrence-discriminating observation, such as the native obligation identity and its verified surface witness. □

Definition 1 (Authority-to-execution instantiation)**.** *An execution trace exhibits authority-to-execution instantiation when an authorization-dependent transition constructs the complete transaction occurrence that is subsequently used to justify a native write, rather than selecting an occurrence that was complete before the authorization event.*

Desired properties. We require: (i) strict opening-before-selection with path exclusion; (ii) at most one live occurrence for a tool/target subject; (iii) affirmative consent and unique reference resolution; (iv) exact-occurrence retirement; (v) cross-source splice freedom; and (vi) no real handler dispatch

without a selection of the same live occurrence. These are safety properties of replay-valid mediated traces under the stated execution contract; the result comes from transition exclusion and guarded dispatch, not temporal precedence alone.

3 HANDLER-OPENED EPOCH SELECTION

3.1 TRUST BOUNDARY AND CANONICAL IDENTITY

The serialized fields in an occurrence are commitments, not the runtime root of trust. At selection and commit, native verifiers re-enter the live runtime objects and require membership and equality with the committed sources. The formal results therefore assume collision resistance of SHA-256 and correct native verifiers, but do not treat a hash by itself as an authority token.

For an admissible parsed JSON-like value x, $\mathcal{C}(x)$ assigns distinct type tags to null, Boolean, integer, finite float, and string values; preserves array order; and sorts string-keyed objects recursively. The outer encoding is compact UTF-8 with no implicit numeric or normalization coercion, so 1, 1.0, "1", and `true` differ while object insertion order does not. Malformed, duplicate-key, non-finite, and unsupported inputs are outside the formal value domain and fail closed before this step.

The proposal artifact q and the user-visible surface certificate s are bound by the trusted proposal-surface hook: s must be generated from the same typed proposal, target, capability, admission, dependency, obligation, and preview witness that form p. The reference resolver observes only the verified surface-event projection and the live set, never model transcript text or an unverified rendering. A compromised renderer that displays a cryptographically valid but misleading surface is outside the current trusted computing base and is evaluated as a deployment assumption rather than hidden by the selector.

3.2 OCCURRENCE IDENTITY

For a surfaced native proposal, EPOCHSELECT records the occurrence

$$p = (k, h_a, h_c, h_m, h_d, o, h_s, h_w), \tag{1}$$

where $k = (\textit{tool}, \textit{target-id})$, h_a is the hash of the typed canonical argument preimage, h_c the native capability certificate, h_m admission, h_d dependencies, $o = (\textit{obligation-id}, \textit{generation}, h_{\text{issued}})$, h_s the proposal-surface certificate, and h_w the deterministic preview witness. The obligation ID, generation, and issued-event hash remain identity fields even when the business arguments are byte-for-byte equal.

Operational core. Let the replay state be $\Sigma = (L, A, C, E)$: live occurrences L, the selected occurrence $A \in L \cup \{\bot\}$, a dispatch-ready committed occurrence $C \in L \cup \{\bot\}$, and an append-only event prefix E. The core transitions are:

$$\frac{\mathsf{VerifyNative}(q) = p \quad L[k(p)] = \varnothing}{(L, A, C, E) \xrightarrow{\mathsf{Open}(p)} (L[k(p) \mapsto p], A, C, E \cdot \mathsf{Open}(p))}, \tag{2}$$

$$\frac{p \in L \quad \mathsf{Consent}(u) \quad \mathsf{Ref}(u, L) = \{p\}}{(L, A, C, E) \xrightarrow{\mathsf{Select}(p,u)} (L, p, \bot, E \cdot \mathsf{Select}(p, u))}, \tag{3}$$

$$\frac{A = p \quad C = \bot \quad \mathsf{Replay}(E) = (L, A, C) \quad \mathsf{SelectedAct}(E, p) = u \quad \mathsf{LiveNative}(p)}{(L, A, C, E) \xrightarrow{\mathsf{Commit}(p,u)} (L, A, p, E \cdot \mathsf{Commit}(p, u))}, \tag{4}$$

$$\frac{C = p \quad \mathsf{Replay}(E) = (L, A, C) \quad \mathsf{LiveNative}(p) \quad \mathsf{H}(\mathcal{C}(a_{\text{final}})) = h_a(p)}{\Sigma \xrightarrow{\mathsf{Dispatch}(p)} ((L \setminus \{p\}, \bot, \bot, E), \mathsf{effect}(a_{\text{final}}))}. \tag{5}$$

$\mathsf{Replay}(E)$ is a deterministic fold from $(\varnothing, \bot, \bot)$ over the typed, hash-valid event grammar in Appendix A; it fails on an invalid source or transition. Thus replay validity is defined independently of the desired invariant. A terminal transition removes its exact occurrence and clears A or C when equal. No other rule emits Open or a native effect; a false premise produces a blocked result without a handler call. Applying this abstract contract to code remains conditional on C1–C5 and the finite refinement evidence below.

3.3 THE ONLY OPENING PATH

The native preview handler first produces artifact q without mutating task state. The mediation hook verifies its capability, admission, dependency, obligation, surface, and witness sources; computes Eq. 1; and appends $\mathsf{Open}(p)$. No other transition can append an opening. In particular, model output, authority observation, commit-time repair, and post-task audit cannot synthesize p.

The replay state maintains $\mathsf{live}(k)$ and the terminal parent of each subject, namely the latest terminal event that ended its preceding occurrence. An opening is legal only if $\mathsf{live}(k) = \varnothing$. If the native lifecycle has advanced to a new obligation occurrence, the old occurrence first receives a terminal invalidation; the child opening names that terminal parent.

3.4 CONSENT AND DETERMINISTIC REFERENCE SELECTION

Authority is the conjunction of two independently checkable predicates. First, $\mathsf{Consent}(u)$ recognizes a frozen set of affirmative forms and rejects negation, questions, reported or quoted speech, conditionals, future promises, and revocation. Second, $\mathsf{Ref}(u, L)$ resolves the utterance against live occurrences L by one of three rules:

1. an exact typed template containing tool, target, and arguments;
2. an explicit target identity named in the authority utterance; or
3. a structural deictic pointer to the unique most recently surfaced live proposal in the runtime's verified surface-event sequence.

Selection is admitted only when consent is positive and the candidate set has cardinality one. Zero or multiple candidates grant no authority and clear all inherited authorization stores. Semantic similarity and model confidence are not authority evidence, and transcript order cannot substitute for a verified surface event.

Studies A–C use this lexical predicate as the sole consent gate. Study D holds occurrence and reference selection outside the task and evaluates classifiers only as offline binary probes; their outputs cannot emit Open or Select or invoke a handler. The frozen binary metric treats only `AFFIRM` as positive; Appendix D prints the implemented lexical intersection and gives parser sensitivity.

Study E freezes a deterministic CONSENT2 repair. Its 14 positive prefixes are selected from SGD training text alone using at least five training affirmations and zero training negations. The safety-marker scope is developed on five STATE-Bench trajectories excluded by exact path from the formal panel, then frozen with the same fail-closed guards. CONSENT2 supplies only the Boolean consent premise. On a positive decision, the typed tool, target, and arguments still come from the surfaced preview; the language predicate does not parse them. To test handler composition without changing reference resolution, the frozen reconstruction uses the trajectory's future confirmed call to preselect one matching preview and exposes it as the sole open occurrence. Study E therefore evaluates binary gating plus exact-once composition conditional on that preselection, not autonomous resolution over the full multi-preview history. Protocol and failure boundary appear in Appendix E.

3.5 COMMIT AND RETIREMENT

Before dispatch, commit verifies the ledger prefix, selected frozen verifier profile, designated baseline or ablation configuration, and the complete runtime proposal snapshot. It recomputes admission and dependency evidence and requires the active obligation ID, generation, issued event (the event that created that obligation generation), proposal surface, and arguments to match p. Any failure terminates before the real handler call. Benign dependency or target drift therefore triggers a generation-coupled re-proposal and a new surface opening; it never silently reuses old authority.

Certified retirement is occurrence-specific. A retirement source must name the opening's tool, target, argument hash, obligation ID, generation, and issued-event hash; that event must occur in the frozen lifecycle prefix (the ordered events replayed for the subject), and the source must contain the matching terminal event and lifecycle tip, the prefix's final event hash. Thus a valid retirement for a later same-key, same-value obligation cannot close the earlier epoch.

3.6 COMPLETE MEDIATION AT THE EFFECT BOUNDARY

Within the abstract scored transition system, Contract C3 represents every effectful commit branch by `_dispatch_native_effect`. The guarded hook performs a final occurrence replay and requires exactly one selected key, matching tool and target identity, a valid opening and typed proposal snapshot, and equality between the snapshot's typed-argument digest and the arguments that reach the hook after any downstream transform. It also checks the live source when that source is still available. A failed check is recorded as a self-hashed dispatch-guard event and returns a `governance_blocked` result before the native handler is invoked. For the repeat-directive strategy, the proposal snapshot is reconstructed from the existing live opening (the *epoch-latched opening*). During isolated revalidation, the adapter exposes only the proposal's business fields (the *semantic preview projection*), so a transient re-rendering is not treated as a new proposal.

Generation refresh uses that same business-field projection. UI-only governance markers and recovery or presentation certificates are excluded from the generation digest, while typed arguments and native result evidence remain included. Consequently, re-rendering an unchanged proposal cannot mint an obligation, but an actual semantic preview change still forces a new generation and surface.

Contract C4 places a final one-shot recheck of the selected key and live proposal snapshot immediately before use. Under that assumption, a bounded interleaving that changes pending arguments, selected identity, or the live source is converted to a guard failure before the effect. This is a conditional transition property, not an empirical concurrency result or a general-purpose lock for asynchronous or multi-writer adapters.

3.7 TRACE PROPERTIES

Let τ be an execution trace whose serialized ledger projection is replay-valid, and let $i(e)$ denote event order. $\mathsf{Dispatch}(p)$ labels the real handler effect at the guarded commit entry; it is not a serialized ledger certificate. Ledger-event verification includes self-hash, previous-event hash, typed source verification, transition legality, and runtime source membership. The following are specification invariants, not a whole-program verification result. Equations 2–5 establish trace consequences assuming Contract C; source refinement must separately show that a concrete adapter satisfies that contract. Serialized certificates themselves are not trusted. Derivations and premise-by-premise scope are in Appendix B.

Invariant 1 (Strict prior opening and path exclusion)**.** *Within the transition system defined above and under Contract C, for every* $\mathsf{Select}(p, u)$ *or* $\mathsf{Commit}(p, u)$ *in the replay-valid ledger, and every* $\mathsf{Dispatch}(p)$ *in* τ*, there exists an* $\mathsf{Open}(p)$ *with strictly smaller position. Moreover, no authority-observation, commit, dispatch, or post-task-audit transition can emit* Open*.*

Invariant 2 (No unselected native dispatch)**.** *Within the defined transition system, every* $\mathsf{Dispatch}(p)$ *whose ledger prefix is replay-valid has a unique earlier* $\mathsf{Select}(p, u)$*, and the live native proposal snapshot at dispatch equals* p*.*

Invariant 3 (Complete mediation for routed adapter writes)**.** *Assume C3–C5, including source-level call-site closure as a premise. Every successful effectful native-handler invocation in the claimed adapter scope is made by the final dispatch hook and is preceded by its final occurrence check. If a downstream argument transform changes the typed arguments, the hook rejects before the handler is called.*

Implementation-to-contract refinement obligation. Invariants 2–3 apply to code only after source-level refinement shows that every claimed write reaches the final guard after its last argument transform. Our 40/40 audit establishes this for two frozen entry roots only. Appendix A.1 maps the abstract state and transitions to the concrete audit targets; C4–C5, other roots, and asynchronous or dynamic dispatch remain open.

Invariant 4 (Per-subject live uniqueness)**.** *Within the defined transition system, after every replay-valid prefix and for every subject* k*,* $|\mathsf{live}(k)| \leq 1$*.*

Lemma 1 (Serialized commitments do not replace live identity)**.** *Assume Contract C and collision resistance of the canonical hash. If an adversary changes any component of an occurrence in a serialized ledger while leaving the corresponding live native source unchanged, the guarded commit rejects before dispatch. Replacing both the serialized component and the live source requires modifying a trusted runtime object and is outside the threat model.*

Lemma 2 (Surface-only reference resolution)**.** *If a user utterance has no exact, target-bound, or uniquely deictic match in the verified live surface projection, then* $\mathsf{Ref}(u, L) = \varnothing$ *and no selection event is emitted, regardless of transcript similarity or model confidence.*

Corollary 1 (Occurrence splice freedom)**.** *Under Contract C, if a valid selection, retirement, or commit consumes sources from two distinct occurrence identities, replay fails before native dispatch.*

4 Conformance and Boundary Evaluation

Studies A–C test finite conformance: completeness (RQ1), baseline separation (RQ2), identity-check necessity (RQ3), benign cases (RQ4), named-shape transfer (RQ5), and in-process restoration (RQ6). Study D tests the language gate against public dialogue labels (RQ7); E tests repaired gating with a preselected action and native handlers (RQ8); F measures full-history input compatibility without target preselection (RQ9). These are development diagnostics, not estimates of natural attack prevalence, deployment safety, legal consent, or end-to-end utility.

4.1 Prospective design and evidence types

Study A is a finite simulation-based deterministic census for one reference adapter over customer-support, travel, and shopping fixtures. Two fixture topologies and four strategies give 24 base contexts; nine substitutions yield 216 cases, plus 24 clean controls, per verifier configuration. It records one development initializer, 20277011. Fixture names, labels, schedule/order blocks, and technical repetitions are not seeds.

The artifact deliberately distinguishes three evidence types:

$$N = (7{\times}240)_{\text{baseline projection}} + (11{\times}240)_{\text{leave-one-check-out projection}} + 72_{\text{benign re-audit}} + 72_{\text{local timing}} = 4{,}464. \tag{6}$$

The first 4,320 rows are deterministic analytic projections over a frozen source-case matrix; 72 benign units re-audit state and exactly-one-call witnesses; only the final 72 units contain newly observed local timings. We never pool these evidence types into a population estimator.

Analytic and benign tickets freeze expected decisions before execution; timing tickets instead freeze correctness criteria, units, order, and repetitions, not a latency outcome. Expected decisions come from a source-distinct deterministic specification; observed rows bind the selected source case, verifier configuration, decision, and row hash. The evaluator enforces

$$N_{\text{plan}} = N_{\text{not emitted}} + N_{\text{inapplicable}} + N_{\text{invalid}} + N_{\text{passed}} + N_{\text{failed}}. \tag{7}$$

A favorable proportion is admitted only when the first three terms and the failure term are zero. This provides a replayable synthetic/self-supervised oracle, not dataset-provided ground truth or human judgment.

4.2 RQ1: completeness and source-distinct recomputation

The unique non-resumed Study A run emits 4,464/4,464 applicable, valid, passing rows with unique cell IDs and valid row hashes. A source-distinct reference checker recomputes the geometry, bindings, semantics, aggregate, and one-shot registry and passes 46/46 predicates. Retry count is zero, `resume_used=false`, and the 375-entry closure reports no missing or drifted source. No model, network, remote resource, or confirmation content is used.

Separate from the estimator, a post-run audit passes 40/40 checks and 40/40 regressions for two frozen primary-study entry roots. This is root-scoped implementation assurance, not experimental rows or a whole-program result.

4.3 RQ2–RQ4: baselines, necessity, and benign boundary

All configurations accept 24/24 clean controls. No guard blocks 0/216 substitutions; full binding blocks 216/216. P0/P2, P3, and P4b protect 24, 72, and 48 cases; the in-process object capability neutralizes all 216. Each of nine identity omissions admits only its matching 24-case slice; the placebo is inert. All 72 benign boundary units pass with matching state and call witnesses. Appendix Tables 5–6 give the fixed-design inventories, not attack or utility rates.

Table 1: **Study E ($N = 128$).** Offline rows isolate the language repair; the native row adds effect witnesses, not accuracy. Labels and preview selection both use the next-action proxy.

Gate and execution mode	TP/TN/FP/FN	Balanced acc.	Positive acc.	FAR
Original rule, offline	10/64/0/54	0.5781	0.1563	0.0000
CONSENT2, offline	25/64/0/39	0.6953	0.3906	0.0000
CONSENT2, native E2E	25/64/0/39	0.6953	0.3906	0.0000

4.4 RQ5: NAMED-SHAPE AND DETERMINISTIC-INITIALIZER TRANSFER

Study B is a separate, frozen real-local-handler extension with a source-distinct rule oracle. Five deterministic initializers, three shapes/domains, two fixtures, and four strategies define 360 contexts with disjoint authority and substitution panels. The one-shot run passes 5,400/5,400; all 1,800 matched non-adapter signatures agree, accepted rows make one state-changing call, and all 3,240 mismatches have no effect. The shapes share one verifier and the initializers are not statistical replications.

4.5 RQ6: IN-PROCESS SERIALIZATION/RESTORE SEPARATION

Study C reconstructs canonical JSON in a fresh in-process store. Across 480 matched three-policy groups (1,440 rows), strict identity blocks all restores, including 120 exact roundtrips; value rehydration accepts all, including 360 nonexact cases; full binding accepts the 120 exact restores and blocks the 360 nonexact cases. Every row satisfies its policy semantics and effect witness. Appendix Table 9 reports the fixed-design counts; OS-process restart and exhaustive persistent-capability comparison are unmeasured.

4.6 RQ7: FALSIFYING THE CONSENT-INTERFACE PREMISE

Study D tests 192 unique, stratified Schema-Guided Dialogue development dialogues (Rastogi et al., 2020). The frozen rule, fixed model, and train-only TF–IDF baseline reach balanced accuracies 0.5000, 0.9844, and 0.9635. The model–rule difference is $+0.4844$ (bootstrap 95% [0.4635, 0.5000]; McNemar $p = 4.85 \times 10^{-27}$), but the model–TF–IDF difference is only $+0.0208$ (95% [-0.0104, 0.0521]; $p = 0.34375$), failing both LLM-specific gates. This falsifies natural-dialogue readiness for the lexical gate; neither learned probe becomes authority. Appendix D gives full counts, sensitivity, protocol, and hashes.

4.7 RQ8: FALSIFYING REPAIRED-GATE COVERAGE UNDER COMPOSITION

Study E composes train-selected, development-scoped CONSENT2 with native handlers on 64 positive and 64 hard-negative next-action proxies from 87 public STATE-Bench train trajectories. It accesses neither model, GPU, network, nor confirmation. Reconstruction uses the future confirmed action to surface one prior preview, testing effect identity rather than full-history resolution.

Against the original offline rule, CONSENT2 has 18 new-only and three old-only correct rows; exact McNemar and a post-hoc trajectory sign-flip test both give $p = 0.0014896$, and trajectory-bootstrap balanced accuracy is [0.6371, 0.7574]. All 64 hard negatives have zero effects (one-sided 95% upper bound 0.0457); each of 25 accepted positives has exact-once selection, commit, handler, dispatch, receipt, and state witnesses. The gain therefore belongs to the language repair; native execution validates composition.

Positive acceptance is only $0.3906 < 0.90$, so the immutable status is `FAIL_GATES`. The threshold is a stringent broad-readiness falsification gate, not a deployment calibration; train acceptance was already 0.1271. With 61/128 multi-preview histories, this is a low-coverage composition diagnostic—not broad C2, natural target parsing, or held-out generalization. Appendix E gives the rejection breakdown, boundary, and no-reuse rule.

4.8 RQ9: FULL-HISTORY COMPATIBILITY WITHOUT TARGET PRESELECTION

Study F processes 104 frozen public ToolTalk and τ^2-Bench files (Farn & Shin, 2023; Barres et al., 2025), containing 10,910 recorded traces in 470 family–domain–task clusters. It retains the complete observed prefix at each user-role event, not a preview chosen using a later action. Six registered read-tool adapters extract candidate descriptors from original tool outputs; a frozen finite grammar examines the original user value. There is no target oracle, selector comparison, model call, or native write.

Among string-valued events, 60,779 fail the authority wrapper and 4,156 fail the query/comparison interface; the remaining 21 parse only bare authority. Table 14 in the appendix reports all marginal counts. The parsed and structurally ready sets are disjoint, so the frozen census stops before any comparison. Appendix F decomposes their intersections from the frozen ledger without rerunning the adapter. Of the 50 structurally ready events, 48 have multiple candidates and nine pass the authority wrapper; none parses. Moreover, 5,188 multi-candidate events have none of the six structural flags yet lack a complete registered catalog. Absence of a flag is therefore not evidence of a usable catalog.

This result is deliberately conditional on an incomplete interface. The adapter retains prior occurrences rather than reconstructing every active lifecycle, permanently marks histories after an unhandled non-read call, and registers only six read tools. The query language is likewise finite. The census therefore does not separate language failure from structural coverage causally, nor establish that an improved resolver would succeed. It supports a concrete boundary: conformance of occurrence binding does not by itself supply an executable full-history language interface.

5 CONCLUSION

EPOCHSELECT makes handlers, not authority, own occurrence identity and revalidates it at synchronous dispatch. Studies A–C exercise this contract; Study D exposes a lexical gap; Study E preserves zero hard-negative effects but fails coverage (39.1% versus 90%) under single-preview preselection. Study F retains full public histories but finds no overlap between the 21 parsed events and 50 structurally ready events under its registered interface. Together these studies support an occurrence-binding contract and measured interface boundaries, not natural target resolution or deployment consent.

AI USE STATEMENT

Generative AI assistants supported development of the conceptual framework and hypotheses, experimental design and synthetic test-case construction, formulation of mathematical claims and proof arguments, method and test implementation, benchmark-data processing, and interpretation of results. They also assisted literature search and synthesis, figure and table preparation, manuscript drafting, editing and formatting, and internal critical review. Checks included automated regression tests, source and ledger audits, and comparison of reported quantities with frozen experimental artifacts. Such checks and AI-generated reviews are not independent human verification or experimental evidence. Quantitative claims in this paper are limited to the frozen Studies A–F artifacts and explicitly scoped analyses; the Study F joint analysis is post hoc. Study D's fixed-model predictions are experimental measurements, not authorization decisions. Responsibility for all submitted content, including AI-assisted material, rests with the authors.

ETHICS STATEMENT

This work uses generated fixtures and existing public benchmark artifacts, without collecting new participant data or commissioning human authorization labels. The measured labels and next-action proxies are not judgments of legal consent. Incorrect binding can authorize an unintended effect, while fail-closed checks can block legitimate work. We therefore state the trusted components and unresolved language-coverage limits explicitly and do not claim deployment safety or recommend treating a successful check as proof of user intent.

REPRODUCIBILITY STATEMENT

Appendices A and B state the transition contract, refinement obligations, and invariant derivations. Appendix C gives the generated-design inventories, and the study-specific appendices document the language, composition, and full-history protocols. Appendix H records artifact identities and recomputation checks. These records support auditing the reported finite designs; they do not constitute an independent confirmation study or imply that a public, anonymized code release has already occurred.

REFERENCES

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

Drew Dean and Alan J. Hu. Fixing races for fun and profit: How to use access(2). In *13th USENIX Security Symposium*, pp. 195–206. USENIX Association, 2004. URL https://www.usenix.org/legacy/events/sec04/tech/dean.html.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *arXiv preprint arXiv:2406.13352*, 2024. URL https://arxiv.org/abs/2406.13352.

Nicholas Farn and Richard Shin. Tooltalk: Evaluating tool-usage in a conversational setting. *arXiv preprint arXiv:2311.10775*, 2023.

Hector Garcia-Molina and Kenneth Salem. Sagas. In *Proceedings of the 1987 ACM SIGMOD International Conference on Management of Data*, pp. 249–259, 1987. doi: 10.1145/38713.38742.

H. T. Kung and John T. Robinson. On optimistic methods for concurrency control. *ACM Transactions on Database Systems*, 6(2):213–226, 1981. doi: 10.1145/319566.319567.

Ben Laurie, Eran Messeri, and Rob Stradling. Certificate transparency version 2.0. Technical Report RFC 9162, Internet Engineering Task Force, 2021. URL https://www.rfc-editor.org/rfc/rfc9162.html.

Xiaoyang Li, Yiqi Wang, Haohui Lu, Zhi Chen, Mo Li, Pingan Song, Mingkai Zheng, and Taotao Cai. MemTX: Transactional belief commit for stateful agent memory. *arXiv preprint arXiv:2607.23929*, 2026. URL https://arxiv.org/abs/2607.23929.

Yuhan Ma and Stefan Schmid. SecureClaw: Clawing back control of LLM agents. *arXiv preprint arXiv:2606.09549*, 2026. URL https://arxiv.org/abs/2606.09549.

Adam Mazzocchetti. Runtime governance for agentic AI: Action-boundary control with trusted provenance and fail-closed execution. *arXiv preprint arXiv:2608.16891*, 2026. doi: 10.5281/zenodo.20262303. URL https://arxiv.org/abs/2608.16891.

Mark S. Miller, James E. Donnelley, and Alan H. Karp. Delegating responsibility in digital systems: Horton's "who done it?". In *2nd USENIX Workshop on Hot Topics in Security*. USENIX Association, 2007. URL https://www.usenix.org/conference/hotsec-07/delegating-responsibility-digital-systems-hortons-who-done-it.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696, April 2020. doi: 10.1609/aaai.v34i05.6394. URL https://doi.org/10.1609/aaai.v34i05.6394.

James Rhodes and George Kang. Proof of execution: Runtime verification for governed AI agent actions. *arXiv preprint arXiv:2607.05397*, 2026. URL https://arxiv.org/abs/2607.05397.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.15817.

Stripe. How Payment Intents and Setup Intents work. Stripe Documentation, n.d. URL https://docs.stripe.com/payments/paymentintents/lifecycle. Online documentation; accessed 2026-09-25.

Uchi Uchibeke. Before the tool call: Deterministic pre-action authorization for autonomous AI agents. *arXiv preprint arXiv:2603.20953*, 2026. URL https://arxiv.org/abs/2603.20953.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024. URL https://arxiv.org/abs/2404.13208.

Zexun Wang. CAVA: Canonical action verification and attestation for runtime governance of agentic AI systems. *arXiv preprint arXiv:2607.13716*, 2026a. URL https://arxiv.org/abs/2607.13716.

Zexun Wang. Proof-carrying agent actions: Model-agnostic runtime governance for heterogeneous agent systems. *arXiv preprint arXiv:2606.04104*, 2026b. URL https://arxiv.org/abs/2606.04104.

Xiaoqi Weng. What you approve is what executes: Consent integrity for black-box LLM agents. *arXiv preprint arXiv:2606.02668*, 2026. URL https://arxiv.org/abs/2606.02668.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. URL https://arxiv.org/abs/2406.12045.

A TRUST BOUNDARY AND TRANSITION CONTRACT

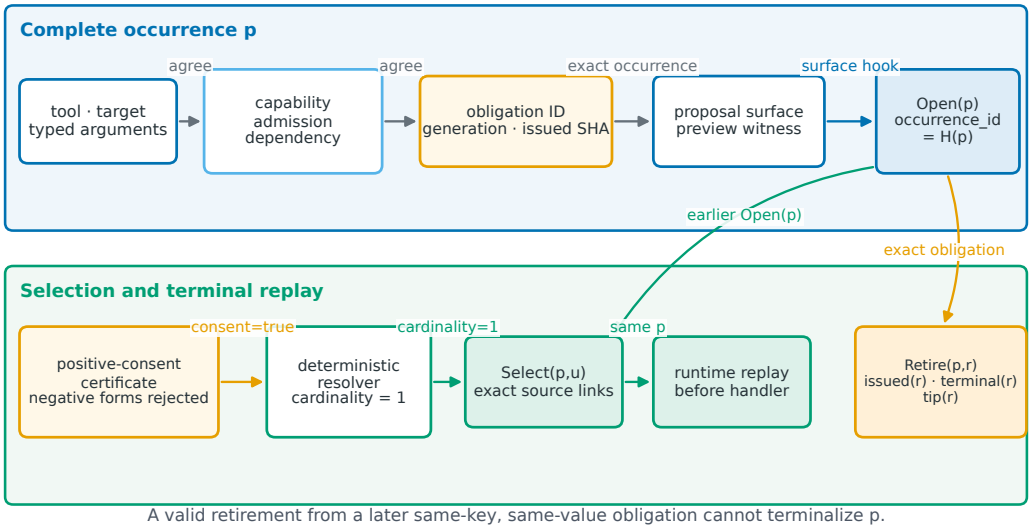

Figure 2: **Occurrence certificate graph.** Selection, retirement, and commit must project to one opening identity. The obligation ID, generation, issued event, terminal event, and lifecycle tip prevent a same-value later occurrence from retiring or authorizing an earlier one.

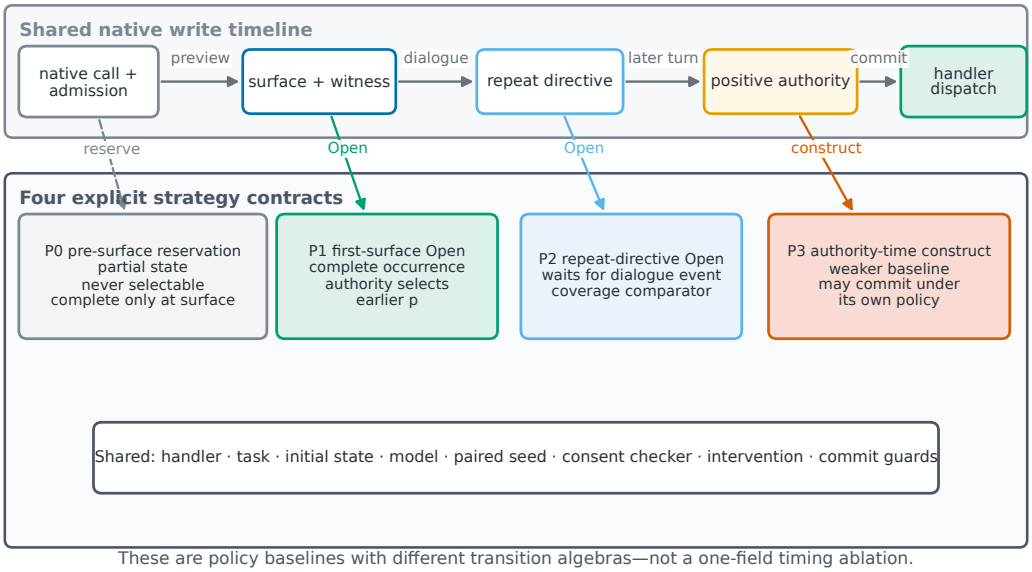

Figure 3: **Four strategy contracts.** All policies share the native handlers, challenge, proposal-boundary intervention, consent predicate, and commit guards, while retaining their own transition algebra.

The trusted execution contract contains typed native verifiers, live runtime objects, the surface hook, ledger replay, and one guarded effectful dispatch premise. Serialized certificates and their unkeyed hashes are adversarial inputs. Contract C assumes final replay and dispatch are synchronous without

an external yield; this connects the ledger's Commit event to the real handler mutation within the invariant model.

Table 2: The refinement boundary is a concrete contract over five trusted components. Serialized certificates are inputs to these components, not members of the trusted computing base.

Trusted component	Assumption checked by the contract	Explicitly out of scope
Typed native verifiers	Capability, admission, dependency, and schema checks are correct and reject unsupported values	A compromised verifier or an undocumented coercion at the parser boundary
Live runtime objects	The selected proposal and effect sources are members of the live runtime and are not forgeable from serialized input	An attacker that mutates the runtime object or the native database outside the adapter
Proposal-surface hook	The user-visible certificate is generated from the exact typed proposal and records one surface event	A trusted renderer that shows a misleading but valid certificate
Ledger replayer	Hash links, source membership, transition legality, and cross-field equalities are checked before commit	A skipped or partially replayed prefix
Guarded effectful dispatch premise	Contract C requires every effect branch to call one final guard; a 40/40 post-run audit discharges the implementation side only for two frozen primary-study entry roots	Another entry root, arbitrary dynamic dispatch, or multi-writer/asynchronous execution without an equivalent lock or versioned atomic primitive

Table 3: Transition ownership in EPOCHSELECT. Authority, commit, and audit have no opening transition.

Operation	Exclusive owner	Required prior state
$\mathsf{Open}(p)$	verified proposal-surface hook	native artifact and complete source bundle; no live occurrence for subject
$\mathsf{Select}(p, u)$	authority observer	positive consent; unique exact/target/deictic reference; prior live $\mathsf{Open}(p)$
$\mathsf{Retire}(p, r)$	lifecycle synchronizer	exact occurrence source; matching terminal event and frozen lifecycle tip
$\mathsf{Commit}(p, u)$	guarded commit path	valid prior selection; current runtime snapshot equals p
$\mathsf{Dispatch}(p)$ (effect)	sole final effect hook	replay-valid commit; final occurrence and typed-argument checks; synchronous unchanged runtime snapshot

A.1 CONCRETE REFINEMENT WITNESS

For the two audited roots, the source audit uses the following abstraction map. L is the set of verifier-valid openings reachable through `_first_preview_open_by_key` or `_epoch_open_by_key`; A is the single key in `authorized_pending_keys_by_turn` together with its verified authority-occurrence oracle; and C is the valid commit-guard result produced inside `commit_pending`. The concrete opening path is `mark_previews_surfaced`, authority observation is `observe_user_turn`, and the abstract dispatch is `_dispatch_native_effect`. The latter rechecks the opening, selection, live snapshot, and post-transform typed arguments immediately before calling `handler(arguments)`. The 40/40 audit checks this map and the absence of a bypass only for those two roots. This is a finite refinement witness, not a proof that all Python paths or future adapters refine the transition system.

Replay grammar and terminal rule. The serialized grammar is $e ::= \mathsf{Open}(p) \mid \mathsf{Select}(p,u) \mid \mathsf{Commit}(p,u) \mid \mathsf{Retire}(p,r) \mid \mathsf{Reject}(b) \mid \mathsf{GuardFail}(b)$, where invalidation and revocation are typed retirement reasons r. Starting from $(\varnothing, \bot, \bot)$, $\mathsf{Replay}(E)$ first verifies each event's self-hash, predecessor hash, typed sources, and cross-field equalities, then folds Open, Select, and Commit by Eqs. 2–4. The function $\mathsf{SelectedAct}(E,p)$ returns the user act in the unique live selection of p and is undefined otherwise. Rejection and guard-failure events clear pending selection and commit state without adding an occurrence. For $x \ominus p = \bot$ when $x = p$ and $x \ominus p = x$ otherwise, the terminal rule is

$$\frac{L[k(p)] = p \quad \mathsf{VerifyTerminal}(p,r,E)}{(L,A,C,E) \xrightarrow{\mathsf{Retire}(p,r)} (L \setminus \{p\}, A \ominus p, C \ominus p, E \cdot \mathsf{Retire}(p,r))}. \tag{8}$$

Any failed premise makes replay undefined. A successful Dispatch is the non-serialized effect and consumes p as shown in Eq. 5; a blocked dispatch instead records GuardFail and makes no native call.

Canonical preimage. The typed preimage used for argument identity is a tagged recursive tree: `null`, `boolean`, `integer`, `float64`, `string`, `array`, or `object`. Integer values are stored as decimal strings, finite floats as hexadecimal IEEE-754 strings, arrays retain order, and object entries are sorted by their string keys. Canonical JSON serialization then uses sorted keys, compact separators, UTF-8, and no ASCII escaping. The adapter begins with parsed values and enforces raw/decoded boundary checks before this preimage is constructed; any application-specific path or numeric normalization remains an upstream schema policy.

B Invariant Derivations

Proposition 1. Fix two executions whose true occurrences differ but whose selector projection and user act are equal. A deterministic selector is a function, so equal inputs yield equal outputs. Requiring acceptance in the live execution and rejection of the stale occurrence demands different outputs and is therefore impossible without an occurrence-discriminating input. The proposition does not prevent an action-envelope system from adding such an input; it identifies the information that its runtime must obtain and verify.

Invariant 1. Induct on the serialized ledger prefix. The empty prefix contains no selection or commit. An opening step can be emitted only by the verified proposal-surface hook. An observation, rejection, or terminal step emits no opening. A selection step verifies an already-live opening with the same complete occurrence identity, hence that opening is earlier. A commit step verifies the selection and recursively the same opening. Finally, Dispatch is not a ledger event: by C3, the sole effectful entry requires that valid commit prefix. Thus neither authority, commit, audit, nor dispatch can synthesize a missing opening.

Invariant 2. At the native mutation entry assumed by C3, the guard verifies one selected occurrence, the frozen policy identity, the complete ledger prefix, and equality between the opening and reconstructed live capability, admission, dependency, obligation, surface, witness, target, and typed arguments. Any coherent rewrite must either fail a typed source verifier, lose live-runtime membership, or violate one of these equalities. The synchronous execution contract admits no mutation between this replay and the effectful handler call, so a dispatched effect has the unique verified selection and unchanged occurrence snapshot.

Invariant 3. Under C3–C5, every effect branch calls the same `_dispatch_native_effect` hook. The guarded transition records a final guard event and invokes the effect only when selected cardinality, tool/target identity, opening validity, and the post-transform typed-argument digest agree with the occurrence snapshot. A changed argument thus returns a governance-blocked result before the effectful call. This is a source-closure argument under C3–C5, not a claim about adapters with additional effect paths or asynchronous hand-offs. For the two named frozen entry roots, the separate source audit checks this routing and final-fallthrough shape directly; C4, C5, helper purity outside the acquired-handler flow, and whole-closure call-site exhaustiveness remain premises.

Lemma 3 (Final recheck detects a visible bounded interleaving)**.** *Assume the final dispatch hook performs its selected-key and live-snapshot recheck immediately before the native effect. If a bounded*

interleaving changes either value after the initial check and before that recheck, the hook returns `governance_blocked` *and makes no effectful handler call.*

Proof sketch. The recheck compares the current selected key and live snapshot with the opening-bound values. A pending-argument, selected-key, or live-source change violates one of these equalities, so the hook records a dispatch-boundary failure and returns before the effect. The lemma is conditional on the recheck being reached and does not establish fairness, mutual exclusion, or correctness for an uninstrumented asynchronous adapter. □

Invariant 4. The live set is empty initially. Only opening adds an occurrence, and its precondition is an empty set for the subject plus the exact terminal parent when one exists. Observation, selection, and rejection add none; invalidation, revocation, retirement, and successful dispatch remove the live occurrence. Every transition therefore preserves cardinality at most one.

Lemma 1. Commit revalidates each typed source and compares its canonical hash, occurrence field, and live membership with the selected opening. A changed serialized component fails its verifier, membership, or identity; a hash-preserving change contradicts collision resistance. Changing the live source itself is outside Contract C.

Lemma 2. The resolver constructs candidates only from the verified surface projection and live set. Unverified transcript text is not an input, and acceptance requires cardinality one.

Occurrence splice freedom. Selection, retirement, and commit each project to every component of Eq. 1. Substituting a source from another occurrence changes the identity, fails live membership, or violates a typed cross-field equality. By Invariant 2, that failure occurs before the only real effect.

C Formal Matrix Inventories

Study naming. Studies A, B, and C in the main text correspond to the content-addressed V113, V114, and V115 evidence chains below; the identifiers are provenance labels, not pooled rounds or additional statistical samples.

The v113 contract freezes 4,464 tickets before execution. Analytic and benign tickets bind their expected decision; timing tickets instead bind correctness criteria, order, and repetitions. Every ticket binds its panel, source case or timing unit, verifier configuration, deterministic development initializer, and source closure. All 24 domain–fixture–strategy base contexts are present. The result satisfies planned=emitted=applicable=valid=passed for all 4,464 cells.

Table 4: Complete v113 denominator reconciliation. Evidence types remain separate even though all rows share the same prospective contract.

Panel	Planned	Passed	Failed	Evidence type
Baseline matrix	1,680	1,680	0	deterministic analytic projection
Leave-one-check-out	2,640	2,640	0	deterministic analytic projection
Benign boundary	72	72	0	source-pair witness re-audit
Counterbalanced timing	72	72	0	observed local measurements
Total	**4,464**	**4,464**	**0**	fixed development design

Table 5: Fixed-matrix baseline separation. Protected means blocked plus neutralized within the enumerated taxonomy; P4a assumes the stated in-process object-identity threat model.

Configuration	Clean	Blocked	Neutralized	Escaped
P1 no guard (null control)	24/24	0/216	0/216	216/216
P0 value only	24/24	24/216	0/216	192/216
P2 strict idempotency body	24/24	24/216	0/216	192/216
P3 tool+args+obligation generation	24/24	72/216	0/216	144/216
P4b message capability+generation	24/24	48/216	0/216	168/216
P4a server-held object capability	24/24	0/216	216/216	0/216
P5 full occurrence binding (ours)	24/24	216/216	0/216	0/216

Table 6: Complete leave-one-check-out inventory. Every semantic omission admits only its matching 24-case diagonal; clean acceptance is 24/24 for every row of the table.

Omitted identity dimension	Blocked	Accepted	Newly admitted slice
Typed arguments	192/216	24/216	typed-argument substitution
Obligation ID	192/216	24/216	obligation-ID substitution
Obligation generation	192/216	24/216	generation substitution
Proposal surface	192/216	24/216	surface substitution
Authority selection	192/216	24/216	selection substitution
Dependency certificate	192/216	24/216	dependency substitution
Admission certificate	192/216	24/216	admission substitution
Preview witness	192/216	24/216	witness substitution
Obligation-issued event	192/216	24/216	issued-event substitution
Full verifier	216/216	0/216	none
Descriptive-label placebo	216/216	0/216	none

The baseline panel has seven configurations over the same 216 substituted and 24 clean cases; leave-one-check-out has the full verifier, nine semantic omissions, and one placebo over the same 240 cases. Contract-ticket `flow` counts bind the benign panel to 24 exact-first, 24 same-value generation-2, and 24 unique-deictic units. Timing-ticket `order_pattern` counts are 36 ABBA/36 BAAB, while `warmups_per_arm_condition` and `technical_repetitions_per_arm_condition` are 5 and 101.

Table 7: Descriptive local timing for the same 72 counterbalanced logical units, each measured under both GC conditions (not 144 independent units). Direct-local unit-median IQRs are [0.009300, 0.018100] ms and [0.009275, 0.018125] ms. Medians of within-unit paired ratios are 1157.758× and 1157.390×; ratios of the two marginal medians are 1037.496× and 1045.316×, respectively.

GC condition	Direct median	Guarded median	Paired-delta median	Guarded IQR
Disabled	0.013300 ms	13.798700 ms	13.788050 ms	[12.148175, 17.297775] ms
Enabled	0.013300 ms	13.902700 ms	13.891750 ms	[12.101650, 17.367050] ms

The OS-process witness records `single_logical_cpu_observed=true`; the unique run has retry count zero and `resume_used=false`. The standard-library reference checker reconstructs all 4,464 tickets and passes 46/46 checks. Table 18 gives the content-addressed execution and analysis chain. These exact counts are a deterministic census over one initializer, not statistical replications.

C.1 V114 NAMED-DESIGN TRANSFER INVENTORY

V114 is a separately frozen study and contributes no row to the v113 denominator. Its five deterministic workload-generator initializer values are `828837942`, `265471229`, `29221287`, `560899624`, and `1353670125`. They are prespecified configuration parameters, not statistical seeds. Three named integration shapes, three domains, two profiles, four strategies, and five initializers give 360 shared contexts. Each context expands into two disjoint panels: six authority forms and nine cross-occurrence substitutions. This is a sum of panel rows, not a 6×9 factorial.

Table 8: Complete v114 denominator reconciliation. The per-shape and per-initializer rows report identical blockwise denominators, not additional pooled observations.

Inventory slice	Rows	Passed	Accepted	Blocked
Authority-form panel	2,160	2,160	720	1,440
Mismatch panel	3,240	3,240	0	3,240
Overall	5,400	5,400	720	4,680
Each named integration shape	1,800	1,800	240	1,560
Each workload-generator initializer	1,080	1,080	144	936

The 1,800 matched non-adapter signatures each contain all three shapes and agree 1,800/1,800. The two authorizing forms contribute 720 accepted rows; the four non-authorizing forms contribute 1,440 blocked rows. Every one of the nine substitution dimensions contributes 360/360 blocked rows. All 720 accepted rows execute exactly one real local handler call and record a nonempty state change; all 4,680 blocked rows execute none and leave state unchanged. The three integration shapes share the same final verifier, so this inventory does not establish independent implementation diversity. The unique formal execution has retry count zero and `resume_used=false`.

C.2 V115 IN-PROCESS SERIALIZATION/RESTORE INVENTORY

V115 is separate from both preceding denominators. It uses the single development seed 20277011 and five deterministic configuration initializers: `2023421333`, `1842590383`, `1525417821`, `1104230286`, and `772155416`. These values enumerate generated configurations and are not statistical seeds. Three domains, two profiles, four strategies, and five initializers define 120 contexts. Four restore cases crossed with three policies yield $120 \times 4 \times 3 = 1{,}440$ rows and 480 matched policy signatures.

Table 9: Complete V115 policy inventory. The target-correct counts are 360, 120, and 480, respectively. Each accepted row has one handler call and state change; each blocked row has neither.

Restore policy	Rows	Accepted	Blocked	False blocks	Invalid execution
Strict ephemeral object identity	480	0	480	120	0
Business-value rehydration	480	480	0	0	360
Full occurrence binding	480	120	360	0	0

The four cases are an exact live roundtrip, rehydration of a complete same-business-value successor, a persisted-surface/successor splice, and a same-occurrence dependency-only drift. Each case contributes 360 rows. All 1,440 rows satisfy their frozen policy semantics; all 480 matched groups pass the realized-workload equality check. The pre-formal smoke is marked ineligible for the estimand. The unique formal execution records 397 source files, retry count zero, `resume_used=false`, and a 16/16 independent reference-check predicate set. Canonical JSON plus a fresh in-process store is the measured boundary; OS-process restart and arbitrary persistent-capability schemes remain untested.

D NATURAL-DIALOGUE CONSENT DIAGNOSTIC

D.1 QUESTION, DATA, AND FROZEN PROTOCOL

Study D (artifact identifier V223) asks a deliberately narrower question than Studies A–C: can the frozen lexical predicate in Eq. 3 recover an official affirmation/negation boundary in adjacent public dialogue turns, and does a fixed language model outperform a strong train-only text baseline? It does not test occurrence construction, reference resolution, dispatch, legal consent, or user utility. Learned systems are offline diagnostic probes and cannot open, select, or dispatch an occurrence.

The source is the Schema-Guided Dialogue corpus (Rastogi et al., 2020), fixed at repository commit `e852981ae34990f4358979625854259302feaa78`. Eligible units pair a system turn that either offers an intent or requests slot confirmation with the adjacent user turn; the official user `AFFIRM`/`NEGATE` act is the gold binary label. The public development split contains 4,352 eligible units before sampling. Three panel seeds `20277021`, `20277022`, and `20277023` define disjoint 64-item panels. Within each panel, the four relation–label strata each contribute 16 units, selected by a prespecified SHA-256 order. Each unit comes from a different dialogue, all 16 V222 pilot dialogues are excluded, and the final 192-unit set covers 13 services. Only train and development files are opened; confirmation/test access is zero.

D.2 COMPARED REPRESENTATIONS

Frozen rule. The same lexical `Consent(u)` implementation used in Studies A–C is applied without adaptation. Only a recognized affirmative form is positive; all other outputs fail closed. Concretely, the base recognizer accepts such patterns as `please proceed`, `go ahead`, `I [explicitly] confirm`, `yes`, `do it`, and first-person `approve`/`authorize`/`accept` or `want to proceed`. The frozen wrapper then requires the same clause to contain `authorize`, `approve`, `confirm`, or the exact phrase `proceed with`, while rejecting questions, conditionals, modal hypotheticals, and reported speech. Thus `I authorize this action` and `please proceed with the exact action shown` survive the intersection, whereas a standalone `yes` or `go ahead` does not. None of the 96 official SGD affirmations survives that intersection, explaining the observed constant-negative rule independently of either learned probe.

Train-only semantic baseline. A class-balanced L2 logistic regression is fit once on 31,360 structurally matched SGD training units. Its fixed input contains intent name and description, relation kind, and system/user utterances, but no gold act. Lowercased word 1–2 grams and character 3–5 grams use sublinear term frequency, smoothed inverse document frequency, and row L2 normalization. After fixed minimum-frequency and feature caps, the realized design has 118,498 features. L-BFGS-B converges in 27 iterations with $C = 1$ and a 0.5 threshold. No development label is used for fitting or tuning.

Fixed model. The local `qwen3.6:27b` model receives the same semantic fields, a frozen prompt and response schema, temperature zero, seed 20277011, and no exposed reasoning trace. The one-shot execution contains exactly one task-neutral readiness call and 192 scored calls, with no retry or resume. Its exact digest is reported with the content-addressed identities below.

D.3 COMPLETE SLICE RESULTS

Table 10: Balanced accuracy by prespecified panel and relation. Panel and relation rows are two decompositions of the same 192 units, not additional observations.

Slice	Units	Fixed model	TF–IDF+LR	Lexical rule	Model–rule
Panel 20277021	64	1.0000	0.9844	0.5000	+0.5000
Panel 20277022	64	0.9688	0.9688	0.5000	+0.4688
Panel 20277023	64	0.9844	0.9375	0.5000	+0.4844
Intent offer	96	1.0000	0.9479	0.5000	+0.5000
Slot confirmation	96	0.9688	0.9792	0.5000	+0.4688
Overall	192	0.9844	0.9635	0.5000	+0.4844

The frozen rule predicts non-affirmation for all 192 units (TP/TN/FP/FN $= 0/96/0/96$). The model yields $94/95/1/2$ and the train-only baseline $90/95/1/6$. The model–rule balanced-accuracy difference is $+0.484375$. A 10,000-replicate dialogue-cluster bootstrap with seed 20277031, stratified over the 12 panel–relation–label cells, gives a 95% percentile interval $[0.463542, 0.500000]$; exact two-sided McNemar $p = 4.8468 \times 10^{-27}$. The model–rule false-accept-rate difference has 95% interval $[0, 0.03125]$. All prespecified rule-coverage gates pass.

Relative to the baseline, the model's balanced-accuracy difference is $+0.020833$. Its 95% interval is $[-0.010417, 0.052083]$, with exact McNemar $p = 0.34375$. It therefore fails both prespecified LLM-specific requirements. Weighting by the four full-development strata (732/567 intent-offer and 2,544/509 slot-confirmation affirmative/negative units) gives descriptive balanced accuracies 0.978894 and 0.975346, a $+0.003549$ difference. This is not a deployment-prevalence estimate.

D.4 PARSER SENSITIVITY AND ERRORS

The model emits 95 `AFFIRM`, 94 `NEGATE`, and three `OTHER` labels. The preregistered binary gate treats only `AFFIRM` as positive, so two positive `OTHER` cases are false blocks and one negative `OTHER` remains a correct non-affirmation. A post-hoc conservative sensitivity analysis counts all three `OTHER` outputs as wrong: the model is then correct on 188/192 units with balanced accuracy 0.979167; its difference from TF–IDF+LR is $+0.015625$, bootstrap 95% interval $[-0.015625, 0.046875]$, and McNemar $p = 0.548828$. The rule-coverage diagnosis remains, while LLM-specific superiority remains unsupported.

All three primary binary errors occur in slot confirmation. Two official affirmations that continue by requesting information are parsed as `OTHER`; one modification of an already confirmed slot is labeled `AFFIRM`. The train-only baseline gets these three cases right. The model recovers several short, elliptical, or cross-domain intent-offer affirmations missed by the baseline, but that advantage does not transfer to slot confirmation.

D.5 CONTENT-ADDRESSED EVIDENCE AND CLAIM BOUNDARY

Table 11: Primary V223 identities. The result also binds the provider and environment journals, baseline coefficients, frozen claim, and review credential.

Bound artifact	SHA256
Fixed-model digest	`a50eda8ed977ab48a12431878896b27ffd5cef552c17af3317d9623b939a7f1e`
Ticket set	`6d5bedc890ddbd470a2be683c85c05a44ee0c9c569d179f0df63fb7fac0c4214`
Scientific source closure	`7c2b85ba63b2448855e30af9fdb346233d0dee4dbf093c7182e8e7700f562a47`
Raw `result.json`	`fe28004c500198091d5d038fec6b884e3066cac060898b1d2d1ee3eba8b8747f`
Result self-hash	`363468535144c21bd74743e070223de9cb8358d8d06d6470644e04977be47996`
Raw model predictions	`03391c71ac86a58561eb8f5caf1e7523fa9bc870d3a96583ff24034674b76362`
Raw TF–IDF predictions	`603c276026f678e3920c76abc69bcf8059d9cae0ef47fea71903d33814653484`

The result records 192/192 prediction rows, 193 provider responses (one readiness plus 192 scored), four provider-environment responses, one baseline fit, retry count zero, `resume_used=false`, and confirmation state `sealed_unopened`. A fresh integrity review reports no P0/P1/P2 finding; a separate result-to-claim review returns a partial verdict. These same-project reviews support artifact integrity and the narrow rule-coverage claim only. Public-corpus contamination, annotation semantics, and the absence of an unpublished external panel prevent a model-specific or submission-level generalization claim.

E STATE-BENCH TRAJECTORY COMPOSITION DIAGNOSTIC

E.1 FROZEN QUESTION, PANEL, AND AUTHORITY PATH

Study E (artifact identifier V226) asks whether a deterministic language-gate repair can compose with the real benchmark-handler path without weakening the occurrence checks. It is not a STATE-Bench leaderboard run. The source repository[1] is frozen at commit `4efcbf2d4fe60df04878859b692d9391f3d5b33a`; the worktree was clean. Five development files used while constructing the protocol are excluded by exact relative path. The formal panel is selected once with seed `20277041` and contains 64 positive and 64 hard-negative turns from 87 train trajectories: 41 customer-support turns and 87 travel turns. Shopping is outside the frozen two-domain sampling scope. No development/test label, confirmation material, model, GPU, or network call is used.

The positive proxy is a user turn followed by an upstream confirmed native action. A hard negative is an authority-lexeme user turn whose next assistant turn contains no confirmed write. These are behavioral next-action proxies, not semantic or legal-consent annotations. At ticket construction, the future confirmed action is used to find a prior preview with the same tool and arguments. The reconstruction surfaces that one preview as the sole open occurrence, applies the candidate consent predicate, and, if accepted, runs selection, commit, and the native benchmark handler. Thus the handler call is real within the local benchmark state, but occurrence disambiguation is controlled rather than measured.

CONSENT2 is frozen before panel execution. From 31,360 matched SGD training units, it retains 14 prefixes having at least five positive examples and zero negatives; prefix selection reads neither SGD development/test nor the formal STATE-Bench panel. Its safety-marker scope was developed using only the five exact-path-excluded trajectories above. The retained forms include `yes please`, `yes go ahead`, `please do`, `go ahead`, `absolutely`, `confirmed`, `correct`, `okay`, `sure`,

[1] `https://github.com/microsoft/STATE-Bench`

and `ok`. The training census is descriptive: positive acceptance rises from 0.0131 to 0.1271 with zero observed training false accepts. The manifest does not retain an extraction preimage digest, so these training counts are not used as a paper generalization result.

The one-shot contract fixes 128 executions, retry zero, no resume, no overwrite, a trajectory-cluster bootstrap with 10,000 replicates and seed `20277042`, and four simultaneous gates: zero false accepts, one-sided 95% false-accept upper bound below 0.05, positive acceptance at least 0.90, and exactly one handler call for every accepted positive with none for every negative. Failure of any gate makes the overall result fail. The 0.90 positive-acceptance threshold is a deliberately stringent falsification gate for broad readiness, not a deployment-calibrated utility requirement. The descriptive 0.1271 SGD-train acceptance already made this gate unlikely; the one-shot panel quantifies rather than conceals that cross-corpus shortfall.

E.2 Results and dependence audit

Table 12: Complete V226 composition counts. Domain rows partition the same 128 turns. Only the E2E row traverses handlers; offline–E2E equality for CONSENT2 is the frozen composition check.

Slice	N	TP	TN	FP	FN	BA
Customer support	41	8	18	0	15	0.6739
Travel	87	17	46	0	24	0.7073
Overall original rule offline	128	10	64	0	54	0.5781
Overall CONSENT2 offline	128	25	64	0	39	0.6953
Overall CONSENT2 E2E	128	25	64	0	39	0.6953

The repaired predicate and end-to-end acceptance agree on all 128 tickets. Every hard negative has zero handler calls, valid dispatches, effect receipts, and state changes. Every accepted positive has one of each and matches the upstream action's tool and arguments. That upstream-match witness is conditional on the construction-time equality used to preselect the preview; it establishes transport integrity, not independent target or argument resolution. The one-sided exact upper bound after 0/64 false accepts is 0.0457297. The balanced-accuracy trajectory-bootstrap 95% percentile interval is [0.637097, 0.757353].

The prespecified comparison with the original offline rule has 18 new-only and three old-only correct rows. Rowwise exact two-sided McNemar gives $p = 0.0014896393$. Because 41 trajectories contribute both labels, we also perform a descriptive post-hoc cluster audit. There are 87 trajectories and at most two sampled turns per trajectory; all 21 discordant rows come from different trajectories. Enumerating all 2^{21} trajectory-level sign flips therefore gives the same exact two-sided $p = 0.0014896393$. This resolves the paired-comparison dependence for the observed discordances, but it does not turn a post-hoc test into a preregistered endpoint.

The failed gate is utility coverage: positive acceptance is $25/64 = 0.390625 < 0.90$. The immutable result status is consequently `FAIL_GATES`. The panel is consumed and cannot be used to tune another rule or be rerun as confirmation. A broader follow-up would require a wholly fresh, trajectory-disjoint, label-blind panel that retains the full preview history and prohibits future-action preselection.

Across the 39 positive misses, six do not trigger the outer authority recognizer and 33 trigger it but fail the frozen affirmative safety predicate. Among hard negatives, the corresponding counts are 56 and five; three more are refused by the repeat native-preview capability rule. The last three cases are the only Study E decisions in which occurrence governance, rather than the language predicate alone, supplies an additional refusal.

E.3 Construct and source-assurance boundary

The source trajectories contain one distinct prior preview for 67 turns and multiple distinct prior previews for 61 turns (27 positives and 34 hard negatives). Study E cannot therefore support a claim that language alone selects the correct occurrence, tool, target, or arguments from a natural history. It

supports only a binary gate plus exact-once native composition given the preselected, singly surfaced preview.

A fresh source-delta audit compares V226 with the passing V225 explicit-taint inventory. It parses 79 current source modules and 11 declared runtime roots, classifies 689 explicit handler-flow events, passes 10/10 checks, and finds the normalized core signature unchanged; every other baseline file remains byte-identical. Its grammar tracks explicit parameters, aliases, transfers, storage escapes, and returns. It does not establish closure over reflection, plugins, native extensions, runtime monkey patching, or arbitrary Python programs.

Table 13: Primary V226 content identities. The contract additionally binds the rule manifest, plan, import origins, runner, handler sources, and tests.

Bound artifact	SHA256
Raw `result.json`	`f72d82200591cc74d14e7e738a91f1b7bcdb044d122c7ba67eb5edc2ee6cf2c6`
Result self-hash	`0f76faa974de44be0569f437664874520 45ecac9559a2cea5e80ec13c4133b4e`
Frozen tickets	`ee56a02d554d135e7d3bd8c689f63467 88546dfad45cce9943da68692b1d69e1`
Source-delta audit	`f70a0d8f1a67dae350a0259b888867f2 e6699b133ea309057f4758c0930ab929`
Construct-scope post-hoc audit	`60528fdd67bc295b844d8af15e1257d0 54a9dabcc0c15d6a76d4b0cfbd244199`

A fresh same-family integrity audit reports artifact `integrity_status=pass` but overall `FAIL` because the primary coverage gate failed and the reconstruction conditions on a future action. A separate result-to-claim review rejects broad C2 and admits only the scoped diagnostic above. Both reviews are provisional rather than independent replication. Confirmation remains `sealed_unopened`.

F Full-History Input Compatibility Census

Table 14: **Study F.** Conditions are parallel marginals, not an attrition funnel. Repeated events and traces are not independent samples. Zero comparison supply is not zero task accuracy.

Input-side condition	Events
All user-role events	110,695
Non-string / string-valued user events	45,739 / 64,956
Multiple candidate occurrences	19,176
Complete registered catalog	80
Structurally ready	50
Parsed query (all bare authority)	21
Parsed query and structurally ready	0
Executable mechanism / identity comparison	0 / 0

Scope and units. V237 is a frozen post-exploratory input census, not a task evaluation. The source inventory contains 104 public files and 10,910 recorded traces. Its ledger retains 110,695 user-role events in 470 family–domain–task clusters: ToolTalk contributes 273 events in 78 clusters; τ^2-Bench contributes 110,422 events in 392 clusters. Repeated trials and configurations remain in the event denominator. Cluster counts indicate distinct tasks exhibiting a condition, not independent statistical replications. All 45,739 non-string user values occur in the τ^2-Bench artifacts and are retained as an explicit unsupported stratum.

What is reconstructed. Only the observed prefix before each user event is available to the adapter. Registered read outputs yield original-text-bound descriptors; no later write argument selects a

preview or constructs a target label. All earlier candidate occurrences are retained. This is not complete active-state reconstruction: unhandled non-read calls permanently set a lifecycle flag, and read tools outside the six-entry registry have no descriptor adapter. These design choices restrict what the census can admit.

Parallel conditions. A complete registered catalog has at least one candidate and binds every candidate through the registered source grammar. Structural readiness additionally requires no lifecycle, typed-profile, registered-schema, builder, duplicate-occurrence, or pending-read flag. Query parsing is a separate condition applied to the original user value. Bare authority does not request the nonidentity evidence needed for a mechanism comparison. The mechanism-evaluable condition requires both structural readiness and the frozen query/evidence prerequisites; its zero count is not a failed target prediction.

F.1 Post-hoc joint precondition analysis

Table 15: V238 event intersections from the frozen V237 ledger. F/T denotes left false, readiness true. Each row partitions all 110,695 events. The corresponding task-cluster counts for F/F, F/T, T/F, T/T are respectively $(470, 5, 14, 0)$, $(470, 5, 392, 1)$, and $(470, 2, 324, 3)$. Clusters can occupy more than one cell and must not be summed.

Left condition × structural readiness	F/F	F/T	T/F	T/T
Query parsed	110,624	50	21	0
Authority wrapper accepted	106,477	41	4,168	9
Multiple candidates	91,517	2	19,128	48

All 50 ready events fail to parse. Nine nevertheless pass the authority wrapper, showing that wrapper acceptance alone does not complete the request interface even on structurally ready prefixes. The 48 multi-candidate ready events occur in just three task clusters. In the multi-candidate stratum, 5,236 events have no structural flag: 48 have a complete catalog and 5,188 do not (three and 230 task clusters, respectively, with overlap allowed). Thus absence of a recorded gap flag is not sufficient for readiness.

Table 16: Exclusive observed structural-flag combinations. Builder, duplicate-occurrence, and pending-read flags are absent in this ledger; the machine-readable analysis retains all 64 combinations in every stratum. Each column is a separate descriptive denominator, not a pipeline stage.

Exact set of flags present	All events	String events	Multi-candidate
None	72,281	42,986	5,236
Registered-schema only	290	290	146
Typed-profile only	1,637	1,637	1,637
Native-lifecycle only	34,960	18,516	10,635
Native-lifecycle and registered-schema	116	116	111
Native-lifecycle and typed-profile	1,411	1,411	1,411
Denominator	110,695	64,956	19,176

The native-lifecycle and typed-profile flags co-occur in 1,411 events; adding their marginal counts would double-count those events. No table estimates how many queries would become resolvable after removing a flag or changing the grammar. These tables identify simultaneous prerequisites, not causal bottlenecks or a counterfactual performance gain.

Evidence and access boundary. The frozen `SUPPLY.json`, `INPUT_LEDGER.json`, and `CLOSEOUT.json` reside under `.aris/analysis/r096-v237-history-supply-20260914-v1/`. The V238 analysis reads only this ledger and its summary, verifies the already recorded ledger digest, and recounts the existing family, domain, and task-cluster marginals. Its joint tables are post hoc descrip-

tions, not a preregistered test or a rerun of V237. No gold label, behavioral target proxy, public selector, task score, model call, GPU run, native handler call, or confirmation access is added. The post-hoc outputs are `BOTTLENECKS.json` and `REPORT.md` under `.aris/analysis/r096-v238-bottleneck-decomposition-20260914-v1/`. The standard-library analysis checks the encoded relationships between candidate counts, catalog completeness, gap flags, readiness, query kind, and evaluability, and reconciles family/domain/task-cluster marginals with `SUPPLY.json` before writing tables. This checks ledger consistency; it does not independently re-extract authority or query semantics from text.

Table 17: Study F and its post-hoc joint analysis. These are SHA-256 digests of file bytes, not the distinct canonical/self-hash fields inside the JSON. The script is `scripts/analyze_r096_v238_bottlenecks.py`.

Artifact	File-byte SHA-256
V237 `SUPPLY.json`	`7a25e7cd8f664b11fdb0a42b32de62fd d491f45d3308003d80d13e5895eda770`
V237 `INPUT_LEDGER.json`	`f4cca6880b0f281bcc6ca8f0a0748a0c 063c185242985b1c1d9166ae1e358192`
V237 `INPUT_FREEZE.json`	`d0d179e05c6cb493ed6412502ee8567a b6fb9c6e6b814358b00823e18d072c98`
V238 `BOTTLENECKS.json`	`776b36d8bf43d5094f1f4f99349daa75 71f3c5890e84e56215ff1e06783e34c1`
V238 analysis script	`6087aadec380a76289a21207fa73e76 016bb20e4403d31965d2de1223c9472c1`

G Confirmation Seal

V113 and V115 explicitly record the single development seed 20277011; V114 has no statistical-seed dimension. V114 and V115 each record five deterministic configuration initializers with no statistical-seed semantics. V223 is a separate supplementary development study with prespecified panel and bootstrap seeds, not confirmation evidence. V226 is a separate failed-gate composition study with prespecified panel and cluster-bootstrap seeds. V237 is an input-only public-source census; its V238 joint analysis does not open labels or add a confirmation run. All six study chains record confirmation status `sealed_unopened`; this manuscript reports no confirmation task, oracle, result, or statistic. A future confirmation run would require a separate prospective authorization and evidence chain.

H Executable Evidence and Reproducibility

The v113 evidence chain is prospective, content-addressed, and single-use. A contract freezes all 4,464 tickets, the sole development initializer, the source closure, expected semantics, and prohibited extrapolations. A freeze receipt binds that contract to the reviewed source snapshot; an OS-process witness and run-attempt record precede scoring. The unique launcher then performs authorization, freeze, run, and reference stages exactly once. The artifact records retry count zero, `resume_used=false`, no model, network, remote server, GPU, or endpoint, and confirmation status `sealed_unopened`.

Table 18: File-level hashes for the frozen v113 execution and analysis.

Bound artifact	SHA256
contract.json	4297b19ebd84db21991983509cd7ec57 5c3318b7a82a4876db2ee34af370b9b8
artifact-core.json	f1ef572bc3025cf3f2e492f89f47f3ab 7846e3cf9afe1222e831a9ad01778124
artifact.json	40c539a49f21200f270073aa2cc3c926 aec0f068d3ffbcc3656d2e57f95da3d2
freeze-receipt.json	8a1b3fe7f341db06ed9d693391d173e0 19aa1b126468888cf7ddd10280007947
run-attempt.json	1bc32d2ac0e3f79a3d386008374ccaad 39d62bd04463447bc071aac0e64e7c3c
os-process-witness.json	39d10a800fe55d054d2e84aa9677687e 2e90379e332e5b448d8a9b675b3306ad
reference-check.json	96547109539fe730dff794c841032111 f187a02c8fd14181c7ff25702f630f48
formal-analysis.json	757563bd77ebd30a21a748c0d80487d bd2b46664009ab059c2a90a6c5012e220

The contract's 375-entry source closure has digest fcbb4070a498d9cfea509275903906f5 0594e0de95e264c21a7000f67da2afb2; the reference checker observes zero missing entries and zero hash drift. Its own file hash is shown in Table 18, and it passes 46/46 checks over the complete geometry, ticket/row hashes, one-shot registry, source identities, aggregate, and seal status. The 4,464 rows have unique cell IDs and unique valid row hashes.

Post-run C3 assurance. A root-scoped source audit passes 40/40 checks; a fresh targeted run passes 40/40 regressions (JUnit SHA256 4fbfc16a1b2e23a39fa77b8abc4629cb 41c9762550d62687c77e0b471d05a8cc). Audit file SHA256: 1359875baf2729f4 552ccc0a76bf84775909d6b9bb6cd3e93faba3c5a1b70b1c; self-hash: da8a5f463bb37b936b55df45d317593219eb47da12de3acb1c09870fbd5ea860. It binds nine sources and a 16-call/one-proxy presence ledger that is non-exhaustive even within the audited subset. This discharges only C3 for the two frozen primary-study entry roots, not arbitrary entry points or a whole-program call graph.

The expected decisions are frozen by a source-distinct deterministic specification. The baseline and leave-one-check-out panels project the same bound 240-case matrix through explicit verifier predicates; they are analytic mechanism evidence, not external attack execution. The benign panel re-audits complete-state and call witnesses from 72 fixed source pairs. The timing panel contains observed local repetitions under counterbalanced order and two GC conditions. This evidence-type separation is serialized in the artifact and is preserved in the analysis rather than hidden behind the aggregate 4,464/4,464 count.

The source closure is a tamper-evident snapshot for this run, not a coverage metric or proof of arbitrary-adapter call-graph dominance. Likewise, the reference checker is source-distinct but remains a same-project deterministic reimplementation; its truth table is synthetic/self-supervised, not external ground truth. The admitted denominator contains only the current v113 rows; no historical row is pooled into it.

Separate v114 transfer chain. V114 has its own prospective contract, source closure, one-shot formal root, reference recomputation, and fixed-count analysis; no v113 row is pooled into it. Its 388-entry source closure has digest 431405efb0d8c0d5db16917ec3e11c9d a1bfb8e6e2980f504b15b95a02559026. The formal run records execution_count=1, retry count zero, and resume_used=false. It emits 5,400 unique ticket/row bindings and the reference checker recomputes all rows, effects, strata, and 1,800 matched non-adapter agreements without mismatch. Table 19 binds the formal and analytical layers.

Table 19: File-level hashes for the separate v114 execution and analysis.

Bound v114 artifact	SHA256
contract.json	fc84ec2f575121cbed199046560407b4a 0f846bd677431a83d42be82288e6d1f
artifact.json	bf52417b894fcb2c363ea1d48b4ac803 efd5d3eaa887b26d789bd65b5cf8cef2
reference-check.json	a4d32398a3522f463400596ccc56a689 81acf150dd146bf7007d71c9b7a34d2e
terminal.json	d8e53525e091910bfe3f21fc1e57bb9 7df54e6c1de877b3a4cab129f98ab2148
formal-analysis.json	a5416711021c8678a083824438cbd335 38db2d2a47ec4d8f0f842c792e74d0f8

The artifact's exact evaluation type is `real_local_handler_execution_with_template_generated_workloads_and_source_distinct_rule_oracle`. Because that oracle and its workloads are generated rather than external ground truth, we conservatively impose a simulation-only claim ceiling; this ceiling is a manuscript scope decision, not the artifact's field value. The three shapes share one final verifier, the five initializers are not statistical seeds, and confirmation status is recorded only as `sealed_unopened`.

Separate v115 in-process serialized-restore chain. V115 has a distinct review credential, a pre-formal smoke marked ineligible, one-shot formal root, source-distinct reference checker, and fixed-count analysis. The smoke is explicitly ineligible for the estimand. The formal closure contains 397 files with digest `5f23dc401b77859cdea5ad653400a2904 f929737e4838d024a6443884a1e787e`. The formal run records one execution, zero retries, no resume, 1,440 unique ticket/row bindings, 480/480 matched realization checks, and no model, network, or GPU use.

Table 20: File-level hashes for the separate v115 execution and analysis.

Bound v115 artifact	SHA256
contract.json	a3473edc5a501a68fd67b687109393b1 401e03c939fb9e679c903f4cf7c4b3f3
artifact.json	61b109ab6c3573f15ae57850dcb933a1 3b37e0d34cf03bacb1cc2562d2ad890e
reference-check.json	fd923b6707789e761fadfc2e95454dcc 60582779a35f0005fa88a00dfb793360
terminal.json	e6b70b5e3f64003b41d99b5a369ef1 bd377d6ec18c029765fdc18e1506224e6c
formal-analysis.json	ff05242d75e6fa374bea031ec3000acc 5292150330cf7087acac9d95fd77e750
experiment-audit.json	5a22d16b627d75e9c5572578ebaa26e7 197140d4bb9b09221478ecfe619bac56
result-to-claim.json	dfd90af84b0bf101070ff28b2e7f43c 0853aa43219654d72623017a9c0a7e24a

Its exact evaluation type is `real_local_handler_execution_with_template_generated_persistence_workloads_and_source_distinct_rule_oracle`. The measured restore boundary is canonical JSON serialization plus a fresh in-process store. The single development seed is 20277011; five configuration initializers are not additional seeds. The evidence therefore supports exact policy tradeoffs within the named family, not operating-system restart, arbitrary persistent capabilities, natural workloads, or population inference.

Artifact availability. An anonymized submission artifact will include the implementation, frozen contracts, ticket manifests, raw row-level results, analysis scripts, and replay/check utilities needed

to reproduce every reported count. Absolute local paths retained in provenance records identify the originating run. The artifact instructions will map their relative equivalents while leaving the content-addressed records byte-for-byte unchanged. For Study F, the bundle will also include the frozen `INPUT_LEDGER.json`, `SUPPLY.json`, and source inventory; the V238 `BOTTLENECKS.json`, analysis script, and its synthetic tests reproduce the post-hoc joint tables. Table 17 identifies these local evidence files. Bundle preparation does not reopen confirmation or convert the census into a task evaluation.

I Related Work

Tool agents and security evaluation. ReAct interleaves reasoning and action (Yao et al., 2023); ToolEmu, AgentDojo, and τ-bench expose tool risk, prompt injection, and stateful agent–user interaction (Ruan et al., 2024; Debenedetti et al., 2024; Yao et al., 2024). Instruction hierarchy addresses model-side priority handling (Wallace et al., 2024). EpochSelect instead constrains the runtime transition that may create a native write occurrence.

Agent memory and effect governance. MemGPT manages long-horizon agent context (Packer et al., 2023), while MemTX applies transactional repair to belief writes (Li et al., 2026). SecureClaw mediates exact requests at preview/commit effect sinks (Ma & Schmid, 2026). Open Agent Passport synchronously applies declarative policy before a tool call (Uchibeke, 2026); Consent Integrity uses a trusted mediator to render the real boundary action and bind approval to execution (Weng, 2026); and Proof-Carrying Agent Actions packages action-open, approval, and closure in a portable envelope (Wang, 2026b). CAVA turns heterogeneous runtime events into canonical action objects with approval-bound fingerprints (Wang, 2026a); Proof of Execution binds authorization, effect, history, and replay in one verifiable execution object (Rhodes & Kang, 2026); and Aegis resolves policy and provenance through a trusted decision layer before effects (Mazzocchetti, 2026). These systems establish policy, canonical action identity, trusted display, or execution evidence. EpochSelect contributes an adapter-level provenance condition for those objects: the native proposal path must create a complete obligation *occurrence* before authority, and selection must resolve to that live occurrence across value-identical generations. An existing system can satisfy this condition by refinement; the claim is not that it cannot add a nonce. Proposition 1 isolates why a value-only projection is insufficient, while the opening rule specifies who may create the missing information and when.

Terminology map. Our vocabulary has direct systems analogues but intentionally narrower semantics. The verified native capability is capability-like designation, yet is not by itself a bearer authorization; the occurrence identifier and surface witness play a nonce-like anti-replay role; the obligation generation is an epoch that separates value-identical retries; the hash-linked ledger is an append-only audit projection rather than the root of trust; an occurrence ID is stronger than a value-only idempotency key because a later identical request is a new generation; and the guarded dispatch recheck is an optimistic validation–use boundary. These mappings locate the construction in familiar mechanisms without claiming novelty for capabilities, nonces, epochs, logs, idempotency, or final validation individually.

Transactions and append-only lineage. Sagas, optimistic validation, and certificate-transparency logs provide compensation, commit-time conflict checks, and append-only lineage (Garcia-Molina & Salem, 1987; Kung & Robinson, 1981; Laurie et al., 2021). EpochSelect uses related mechanisms; object capabilities likewise couple unforgeable designation with narrow authority (Miller et al., 2007). Production APIs such as PaymentIntent also create a durable transaction object before a later confirmation (Stripe, n.d.). We make this pattern explicit for model-mediated native writes and add deterministic reference selection, typed source membership, and cross-generation splice rejection. The contribution is therefore transition ownership plus occurrence-sensitive replay, not the existence of a hash chain, capability, or epoch.

Table 21: Closest abstractions. Rows describe their stated primary binding, not limits on extension. EPOCHSELECT is an adapter refinement for native occurrence origin and generation binding, not a replacement for policy, canonical action objects, trusted rendering, or execution receipts.

System	Primary binding	Distinction from EPOCHSELECT
OAP (Uchibeke, 2026)	Policy decision to an intercepted tool call	No native preview-occurrence generation semantics
Consent Integrity (Weng, 2026)	Trusted rendering of the real boundary action to exact execution	Targets approval-channel spoofing, not same-value lifecycle occurrence reuse
PCAA (Wang, 2026b)	Portable action-open/approval/outcome certificate	Runtime-neutral action envelope rather than handler-owned native occurrence
CAVA (Wang, 2026a)	Canonical action object, fingerprint, approval receipt	Canonical identity does not by itself establish native lifecycle origin before authority
Proof of Execution (Rhodes & Kang, 2026)	Contract, causal event stream, and replay context	Verifies governed execution; occurrence minting remains an adapter refinement obligation
Aegis (Mazzocchetti, 2026)	Trusted policy/provenance decision before effect	Controls action admission rather than value-identical native occurrence generations
PaymentIntent (Stripe, n.d.)	Payment object created and tracked before confirmation	Domain state machine, without agent reference resolution or source-splice proof
EPOCHSELECT	Native handler occurrence to later selection and dispatch	Exact obligation generation, live-source replay, and deictic occurrence selection

J SCOPE AND LIMITATIONS

Conditional system boundary. The specification invariants depend on typed native verifiers, live runtime membership, a faithful proposal surface, complete ledger replay, and source-level closure of the guarded effectful entry. The primary-study source audit discharges the implementation side of C3 only for its frozen full-verifier and timing roots. Its root-external ledger is presence-checked but non-exhaustive even within the nine audited source files; it is not a whole-program control-flow proof or coverage of arbitrary dynamic dispatch. An undocumented effect path, compromised trusted component, misleading trusted rendering, hash collision, or external mutation of the live runtime is outside Contract C. An asynchronous or multi-writer deployment needs an equivalent lock or versioned atomic primitive; the conditional final recheck is not a general concurrency proof. Application-specific Unicode, path, numeric, and schema normalization remain upstream obligations.

Finite generated-design evaluation. Studies A–C are censuses of separate prospectively frozen deterministic development designs, not samples from natural model or user behavior. Study A contains 4,320 analytic verifier projections, 72 benign source-pair re-audits, and 72 local timing units for one reference adapter. Study B contains 5,400 real-local-handler rows from two disjoint template-generated panels over three named integration shapes and five deterministic workload-generator initializers. The shapes share one final verifier and are not independent implementations; the initializers are configuration parameters, not statistical seeds or independent samples. Study C contains 1,440 real-local-handler policy rows over 480 matched generated workloads. Its expected policy semantics and target decisions are source-distinct and replayable but not dataset-provided ground truth, human judgment, or a third-party implementation. The only development seed recorded in Studies A–C is 20277011. The five initializer values in Studies B–C provide no additional statistical replication. Their exact fixed-design counts support no variance, significance, attack-prevalence, or population claim.

Natural-language diagnostic. Study D is not a natural-user or legal-consent study. It is a balanced, stratified sample of 192 public SGD development dialogues, labeled by two official dialogue-act categories rather than human judgments of authorization. Its equal relation–label allocation is

intentionally different from corpus and deployment prevalence; the full-development reweighting is descriptive only. The public corpus may have appeared in model pretraining. We test one fixed model configuration, one train-only supervised baseline, and one frozen lexical rule on a single corpus family. The three model `OTHER` outputs also make the primary score a binary gate metric, not strict three-class accuracy. Same-project integrity and claim reviews are provisional assurance rather than independent replication. No unpublished external panel, cross-dataset test, human authorization annotation, or independently implemented evaluator is available.

Trajectory-derived composition diagnostic. Study E uses 128 action-proxy turns from 87 STATE-Bench train trajectories in only customer-support and travel; it is not the official held-out benchmark or an independent legal-consent annotation. Forty-one trajectories contribute both labels, although every paired discordance belongs to a distinct trajectory and the post-hoc cluster exact test matches the rowwise McNemar result. More importantly, the reconstruction uses the future confirmed action to choose one matching preview before execution; 61/128 original histories contain multiple distinct previews. The study therefore tests consent classification and exact-once handler composition given a preselected occurrence, not natural occurrence, tool, target, or argument resolution. Its 0/64 hard negatives have a 4.57% one-sided upper bound, while 25/64 positive acceptance fails the 90% gate. The consumed panel cannot be tuned on or rerun as confirmation.

Full-history compatibility census. Study F is an input-side census of frozen public artifacts, not a new held-out test. Its 110,695 user-role events include 45,739 non-string values; events and recorded traces can share a task cluster. Six registered read tools and a conservative persistent lifecycle flag restrict structural readiness; a finite grammar restricts parsing. The post-hoc joint tables describe co-occurring barriers and cannot attribute causes or estimate a counterfactual repair gain. Zero executable comparisons does not imply zero target supply in the original benchmarks. No labels, target accuracy, population confidence interval, or confirmation result is obtained.

Baseline semantics. The weaker baselines are mechanism projections over the same 216 substituted and 24 clean cases, not claims about complete deployed products. In particular, the server-held object-capability baseline neutralizes all enumerated substitutions only while its object identity remains live in one process. Study C relaxes that assumption only through canonical JSON serialization and a fresh in-process store. Strict identity then over-blocks exact roundtrips, whereas value-based rehydration can rebind authority to a nonexact occurrence. This is not an operating-system process restart or an exhaustive comparison of persistent-capability designs. Non-monotone protection counts reflect which frozen identity fields each baseline binds, not a general system ranking.

Leave-one-check-out scope. Each semantic omission admits exactly the corresponding prespecified 24-case slice, while the descriptive-label placebo is inert. This supports necessity of all nine checks for complete coverage of this matrix and implementation. It does not prove that the taxonomy is exhaustive, that each predicate is sufficient in isolation, or that arbitrary adapters expose the same boundary.

Benign and performance endpoints. Passing 72/72 benign units shows boundary conformance for the enumerated derivations; it is not a natural-workload false-positive or utility estimate. Timing is handler-relative and single-machine. The guarded arm includes the full Python replay path, whereas the direct arm is a minimal local handler call. ABBA/BAAB ordering and two GC conditions address fixed-order sensitivity, but 101 within-unit repetitions are technical measurements, not independent units or seeds. We make no causal overhead, end-to-end latency, non-inferiority, or deployment performance claim.

Unmeasured endpoints. Studies A–C, E, and F use no model, network, remote server, GPU, or user-utility endpoint. Study D uses a fixed local model and GPU only as a non-authoritative offline classifier; it does not execute handlers or measure end-to-end utility. Confirmation remains sealed and unopened, and no confirmation statistic is reported. Study B reduces only the single-shape and single-workload-initializer ceilings within its generated family; Study C reduces only the in-process serialization/restore ambiguity; Study D establishes only a public-development rule-coverage defect. A statistically replicated, independently implemented, externally held-out, asynchronous/concurrent, utility, deployment-latency, business-oracle, security, legal-consent, or confirmation claim requires

a separate prospective study. Study E's source-complete audit is syntactic under a frozen explicit taint grammar; it does not cover reflection, plugins, native extensions, monkey patching, or arbitrary Python programs.

Operational impact and human factors. Fail-closed replay can over-block legitimate work when upstream evidence is missing, stale, or mis-normalized. Conversely, a green runtime decision can encourage automation complacency if operators mistake conformance to this local contract for correct user intent or safe deployment. Production use therefore needs observable recovery, escalation, and human-facing outcome evaluation; none is measured here.