Inference, Authorisation and Consequential Execution: A Complete-Contract Subsumption Test

PAPER-0 Public Manuscript v0.3.1 - released via Weidu Fanling Official Manuscript Page
Correspondence identity: Weidu Fanling
Original manuscript date: 8 September 2026

Cold review date: 9 September 2026

Controlled revision date: 9 September 2026

Publication-preparation date: 10 September 2026; bibliography frozen 10 September 2026; public release date: 10 September 2026

Project state: ROOT-0 / K0/CHI; r8 frozen and unchanged. Status: Public manuscript v0.3.1; not peer reviewed or endorsed.

Abstract

This study asks whether a specified consequential-action contract implemented in the frozen K0/CHI r8 demonstrator requires an enforceable distinction beyond a competent composition of established mechanisms. The comparison was prospectively preregistered and executed as a deterministic finite conformance study over synthetic authenticated records, mutable policy state, a controlled clock, faults, a controllable plant, and independently stated observations. Arm A was actual frozen r8 behind the declared thin semantic adapter. Arm B was a separately implemented conventional composition of rooted authorisation, stateful policy enforcement, atomic check-and-consume operations, scoped one-use grants, occurrence-bound evidence, finite response release control, and governed compensation/remedy workflow.

Across the 57 input records designated comparable by the preregistration, A and B matched on the declared supported projection, and both matched the P1 oracle; the same held for all 308 comparable prefixes. Twenty-six positive-completion checks passed per arm. All 12 sensitivity inputs for ten named mutants produced their preregistered bad observations. Four full-profile diagnostics were not comparable as full r8 guarantees: they preserve the predeclared absence of arbitrary response confidentiality enforcement and automatic timed remedy in actual r8. Consequently, the study does not establish full 12-family conformance of r8.

The permitted substantive conclusion is therefore bounded: under the supplied semantics and controlled environment, the tested r8 distinctions were reproduced by the declared conventional mechanisms, and no source-grounded irreducible enforceable distinction survived this corpus. The preregistered stopping rule closes the distinctive-architecture escalation branch for this tested scope. The result is not a theorem of universal equivalence, a production-safety claim, a claim of institutional legitimacy, or evidence of an engineering cost advantage. Same-investigator design, shared fixture and oracle infrastructure, fixed-setup witness minimality, and a transparently post-outcome recording pass remain material limitations.

Keywords: runtime authorisation; consequential action; usage control; capability security; information-flow control; obligations; compensation; conformance testing; AI agents

1. Introduction

Tool-using and agentic systems increasingly separate model inference from effectful execution. A model may propose an action, a runtime may decide whether the action is authorised, an actuator may commit or partly commit an external change, and a separate observation path may determine what actually happened. These distinctions are not new as individual ideas. Protected mediation, proof-carrying requests, stateful usage control, information-flow enforcement, obligation models, compensation, and runtime agent governance all provide established mechanisms for parts of this problem [1-8, 10-19].

K0/CHI r8 was developed as a compact demonstrator that makes several consequential-action distinctions explicit at one boundary: admissible evidence is not itself standing effect authority; authority and permissions are versioned and scoped; obligations can be occurrence-bound; purpose-limited grants can be consumed; and permit admission is distinct from later emission. ROOT-0 was opened to test a stronger question than whether those distinctions are useful. The question was whether, when a conventional comparator is given the same trusted facts, policy, clock, state, observations, and control surface, any enforceable property of actual r8 resists competent conventional construction.

That question matters because an integration profile can be useful without constituting a new primitive. Conversely, a system can contain a genuinely useful architectural boundary even when every local mechanism has precedent. The appropriate test is therefore not vocabulary comparison. It is a constructive null: build the strongest conventional composition warranted by prior art and compare it with the frozen reference under a prospectively fixed contract.

PAPER-0 reports the resulting E1 v0.1 experiment. The principal finding is negative with respect to distinctive architecture but positive with respect to reproducible engineering structure. On every record and prefix that the preregistration classified as comparable, the conventional composition reproduced the supported r8 projection. The finite corpus also retained two explicit gaps in actual r8: arbitrary response confidentiality and automatic timed remedy. Those gaps prevent a full-profile conformance claim and, importantly, were not repaired by hiding them behind the adapter.

The contribution of this paper is therefore methodological and empirical rather than a novelty claim about an invented mechanism: (i) a complete-contract comparison in which every requested family receives an explicit support disposition; (ii) a frozen A/B construction checked against an independently stated oracle; (iii) positive-completion and sensitivity controls intended to detect a vacuous deny-all comparator; (iv) full-profile diagnostics that preserve unsupported reference properties rather than silently filling them; and (v) a stopping decision that treats bounded subsumption as a reason to stop escalating a distinctive-architecture hypothesis for the tested scope.

1.1 Research question

The primary question is:

> Given equal information, policy, clocks, state, observation and control, what observable property of actual frozen r8, if any, resists a competent conventional construction?

The study does not ask whether the K0/CHI vocabulary is new, whether its integration profile is useful, or whether one implementation is easier to deploy. Those are separate questions.

1.2 Non-claims

This manuscript does not claim that language models cannot abstain, that softmax forces consequential action, that proposal/execution separation is novel, or that a hidden or culpably concealed defect has been identified. It does not infer that specialist outreach, silence, or agreement constitutes peer review or endorsement. It does not equate a policy decision with institutional legitimacy, a permit with an observed outcome, or an obligation record with delivered remedy.

The term complete-contract is used narrowly: the experimental contract assigns every requested family a declared support/comparability disposition. It does not mean that actual r8 implements every full-profile property, and it does not mean that E1 proves complete refinement of the r8 codebase.

2. Terminology and prior art

To avoid novelty by renaming, this manuscript uses established operational terminology wherever possible.

Working distinction	Operational term used here
Inference	Claim or proposed action
Evidential standing	Admissibility, sufficiency and integrity of decision evidence under a stated rule
Authority	Rooted competence or delegation over a declared scope
Permission	Current policy decision for an exact operation
Consequence	Separately identified dispatch, commit and recipient-visible outcome
Return/remedy	Cancellation, governed compensation, notice, correction and duty fulfilment, treated as distinct operations

These distinctions are analytical, not a mandatory serial pipeline. "Independence" is also disambiguated: investigator independence, authenticated actor capability, process isolation, statistical independence, and institutional competence are different properties.

2.1 Conventional security and authorisation mechanisms

Reference-monitor reasoning and complete mediation have long framed protected boundaries between requests and effects [1]. Proof-Carrying Authentication makes authorisation contingent on a relying party checking a request and proof under trusted policy rather than treating the mere existence of information as authority [2]. UCON explicitly models authorisations, obligations, conditions, mutable attributes, and ongoing decisions [3]. XACML similarly separates policy decision from enforcement and can carry obligations/advice, although a policy representation does not itself execute a physical remedy [4].

For information release, relational information-flow work on attacker control, integrity, declassification, and endorsement supplies vocabulary for reasoning about whether high/private information can influence a public trace [5]. Schneider's execution-monitor framework is a reminder that enforcement claims must be scoped to the class of properties and control available to the monitor [6]. ODRL distinguishes duties, consequences, and remedies at the policy-representation level [7]. Sagas supplies established compensation structure for long-lived transactions, while also making clear why compensation need not literally undo every real-world effect [8].

The reject option predates modern language models by decades. Chow's recognition/reject trade-off formalises explicit rejection as a decision option [9]. More recently, the Scientist AI Predictor explicitly includes a deployed noprediction possibility and separately notes cases where abstention itself can carry risk [12]. These sources are enough to retire any premise that consequential architectures require inventing refusal as a new primitive.

2.2 Contemporary agent-runtime neighbours

Several recent systems make the proposal/execution boundary explicit. SARA separates action induction from runtime authorisation and keeps user-rooted authority distinct from action-origin history [10]. PACT enforces argument-level provenance contracts and conditions its guarantees on correct provenance and contract semantics [11]. CaMeL separates trusted control flow from untrusted data flow and enforces capability-based information-flow policy around tool use [13].

AADP specifies per-action authorisation over mutable state, budgets, reservations, approval lifecycle, obligations, and idempotency; the cited -02 document is an individual Internet-Draft updated 1 September 2026 and is not an IETF standard or endorsement [14]. AAE specifies a machine-evaluable authorisation envelope around mandate, constraints, and validity; its -02 draft is dated 6 September 2026 and is likewise work in progress [15]. PCAA organises runtime governance around proof-carrying action certificates and outcome closure [16]. Proof of Execution separates planning, enforcement, effect, and recordkeeping under explicit deployment assumptions [17]. Aegis treats model outputs as proposals and enforces a trusted action boundary before effect [18]. CPEX documents a deterministic policy/reference-monitor composition including authorisation, delegation, redaction, information-flow checks, and audit; in this study its published documentation is related work, not an executed arm [19].

These neighbours are used to locate mechanisms and terminology, not to claim identical semantics or to award any unrun system a score. SARA, AADP, AAE, PCAA, Proof of Execution, Aegis, CPEX, CaMeL, and PACT remain UNRUN in E1 v0.1.

2.3 Provenance of the question

The question that became ROOT-0 arose through an earlier QGIN -> QI/K0/CHI -> backtick line of exploration. That history explains why consequential authorisation and a visible boundary marker became salient, but it is nonevidentiary for the empirical result reported here. PAPER-0 does not rely on the broader QGIN archive to establish technical priority, correctness, or causal support.

3. Frozen study identity and design

E1 was governed by a preregistration frozen before execution. The experiment and its source/results record were not modified for this manuscript stage.

3.1 Integrity pins

Artifact	SHA-256
E1_Complete_Contract_Preregistration_v0.1.md	24da85c0692025ffe5f7d2fd464ba78e7be71ce337e533097aa6f5f41b87d316
e1/PRE_EXECUTION_FREEZE.json	e56a634be82f3e941d95f8d0b0254b77dfb6315ac90c11beddbd3454683b1f85
E1_RESULT_AND_DECISION_RECORD_v0.1.md	cba90a53f735f50cc7101257bd492a82ac91f09117eb1776ca3c16311f8423c8
frozen r8 reference_impl/k0chi.py	65d6612749fd43de9cb13a893a845c766dfd4b143a97f5bf244d9e596bc695a3

The Stage 2 package manifest also pins the nested Stage 1 ZIP as 79f1a44787554784563be9c42ce61373ecb9ef0dd2ccf98685d860d98d5f0737 and the original handoff ZIP as a5e018b1877c0f67ca9eb844ffcb784f7c175078b6ce65e88b3f05419b5a83cf. Stage 1 had independently reconciled the original handoff, its frozen r8 archive (2e82051c77efee85a5b4386585ff7f92bf47509cf804559c249bc0f45e0efcc3), the r8 internal manifest, and the reference implementation pin above.

The code/input/oracle/support freeze was written before the first A/B instance. The materialisation gate checked input hashes, syntax, finite expansion, and oracle tables before outcome collection. A single pre-outcome correction in B replaced fixture-flag-based copy rejection with actual opaque-token registry membership. This correction implemented a component already required by the frozen contract; it did not change a scenario, criterion, support label, oracle table, or frozen r8 source.

3.2 Arm A: actual r8

Arm A was actual verified r8 behind the preregistered thin semantic adapter. The adapter translated authenticated synthetic domain records into the reference API but was not allowed to supply missing native guarantees. In particular, it could not install an arbitrary response-confidentiality filter and call that an r8 guarantee, nor could it install an automatic timed remedy worker into r8.

Source inspection before E1 established several implemented properties relevant to the comparison: effect authority and evidence are distinct; currentness is class-sensitive; occurrence discharge is tied to a live obligation occurrence; purpose-read grants are scoped and consumable; permit admission and emission are separate; and response emission checks current response authority and observation contract. The same inspection also established two relevant absences. emit_response does not inspect arbitrary candidate text for protected content, and the reference API lacks an automatic clock-driven remedy lifecycle with outcome collection, restoration, notice, and deadline handling.

3.3 Arm B: conventional composition

Arm B was separately implemented by the same AI-assisted investigation. It did not import r8 or the oracle. Its declared components were conventional: record validation; exact rooted delegation; mutable consent/approval policy; atomic check-and-consume; a shared quota; occurrence certificates; narrow one-use summary grants; a finite response release policy; and a governed remedy worker.

B was deliberately made strong. The null under test was not "can a weak access-control wrapper match r8?" but whether a competent composition of known mechanisms could reproduce the supported projection under identical supplied semantics. Weakening B would have made a difference easier to find but less informative.

3.4 Shared environment and supplied semantics

Both arms received the same synthetic authenticated records, policy, clock, scheduled events, fault injection, controllable plant, and observations. No language model or live service supplied candidate actions. The experiment therefore tests runtime conformance under a controlled semantic substrate; it does not test natural-language policy interpretation, model intent inference, real credential compromise, or uncontrolled external channels.

The use of supplied semantics is load-bearing. A policy engine can only enforce the meanings represented in its inputs and contracts. E1 does not show that real organisations would agree with the synthetic authority, consent, purpose, or remedy rules. It shows what followed when both arms were given the same rules.

3.5 Comparison relation, coverage, and stopping rule

There were 61 named primary input records across 12 scenario families. Some records deliberately overlap. Fifty-seven were designated comparable for the supported r8 projection; four full-profile diagnostics retained known unsupported reference coverage. For every comparable fixture and prefix, the declared projection of A and B was compared under identical authenticated inputs and checked against the independently stated P1 oracle. This structure prevents "A equals B" from counting as success if both are wrong against the oracle.

Positive-completion checks were required so that a comparator could not pass by denying every consequential operation. Sensitivity mutants were also preregistered. A mutant had to produce its named bad observation rather than merely flip an internal flag.

The stopping rule was substantive. If the strong conventional composition reproduced the supported projection and no defensible irreducible residual survived the causal checks, escalation of the distinctive-architecture branch would stop for that scope. A positive result for subsumption was therefore a stopping result, not an invitation to rename the same integration profile as a new primitive.

4. Results

4.1 Primary accounting

The completed run recorded the following finite observations.

Observation	Result	Interpretation
Primary input records completed	61 / 61	Named records across 12 families
Primary arm replays	122	61 per arm
Scheduled post-setup events	272 per arm	Controlled schedule, not arbitrary race enumeration
Recorded prefix observations	333 paired / 666 arm	Includes each fixture's initial state
Comparable records	57 / 57 matched	A/B equality and P1 oracle conformance
Comparable prefixes	308 / 308 matched	Exact supported projection after every prefix
B against full-profile oracle	61 / 61 records, all 333 prefixes	Includes four full-profile diagnostics
Positive completion goal checks	26 / 26 per arm	52 checks total
False denials on comparable expected-allow events	0 per arm	Expected revocation/quota denials excluded
Unauthorised ordinary tool dispatches	0 per arm	Separate from response-disclosure diagnostics
Sensitivity input records	12 / 12 detected	Ten named mutants; M3 and M9 use two inputs each
Unexpected primary mismatches / exceptions	0	No patched primary rerun

The key empirical result is therefore bounded subsumption of the supported projection under supplied semantics: all 57 comparable records and all 308 comparable prefixes matched between A and B and conformed to the preregistered oracle. These are finite conformance observations. They are not independent statistical samples, confidence intervals, or a theorem over every possible r8 trace.

4.2 Useful completion was preserved

Each arm passed all 26 positive-completion goal checks on comparable inputs. This matters because a fail-closed deny-all construction could otherwise mimic many safety properties while being operationally useless. In the same comparable corpus there were no false denials of expected-allow events and no unauthorised ordinary tool dispatches in either arm. Expected denials after quota exhaustion or revocation were not counted as false denials.

The result therefore cannot be summarised merely as "both systems blocked bad actions." Under the tested fixtures, both also completed the operations that the contract expected to be allowed.

4.3 Full-profile diagnostic S11: response confidentiality

S11 intentionally extends beyond a guarantee implemented by actual r8. With current response authority and observation contract in place, A accepted candidate response text containing either REFUSAL:protected=0 or REFUSAL:protected=1, then later emitted the safe template. Because the protected bit differed across the paired high/private inputs, A's public traces differed. Secret-release cases were 2/2 for A.

B's finite release rule rejected both secret-bearing candidates and emitted only REQUEST_NOT_COMPLETED, leaving the paired public traces equal. Secret-release cases were 0/2 for B.

The shortest source-grounded separating event is one response attempt under the fixed setup. This is not a newly discovered breach of an r8 confidentiality promise. The promise was absent from the reference implementation before E1. The diagnostic therefore remains category 9, unsupported coverage. It identifies an output-enforcement component that a fuller profile would need; it does not establish a new K0/CHI primitive, and B's finite template rule is not a theorem of natural-language noninterference.

4.4 Full-profile diagnostic S12: timed remedy

S12 similarly preserves a full-profile property not supplied by actual r8. In both arms the dispatched operation produced the stipulated partial external state (UNASSIGNED,2) with an ERROR acknowledgement. The partial state is recorded as an actual partial commit rather than being erased from the narrative by the failed acknowledgement.

A exposed no automatic worker/duty lifecycle that would restore the earlier (10,3) state, collect an independent outcome witness, notify the recipient, and meet a deadline. The partial state therefore remained.

B created the declared remedy duty. In the service-available fixture, service at t=1 restored (10,3), an independent witness was obtained, CHANGE_FAILED_RESTORED was delivered, and the duty was discharged before the t=2 deadline. In the no-service fixture, B did not fabricate success: the partial state remained and the duty was explicitly reported FAILED_OVERDUE at t=3. Neither arm was credited with completing the originally requested move to slot 11 in either failure fixture.

The first separating prefix is issue followed by failed dispatch, with shortest fixed-setup event length two. The difference belongs to workflow/actuator integration under stipulated ROOT competence and service assumptions. It is not evidence that the synthetic authority rule is institutionally legitimate. As with S11, the correct primary classification is unsupported reference coverage, not an architectural victory over an implemented r8 guarantee.

4.5 Sensitivity controls and witnesses

All ten named mutants were detected on their 12 designated inputs.

Mutant	Input	Intended lost property	Shortest event length
M1	S02.main	Authority/admissibility distinction	2
M2	S03.main	Authority/admissibility distinction	2
M3	S04.main	Revocation discipline	3
M3	S09.main	Revocation discipline	3
M4	S06.substitute	Exact action binding	3
M5	S07.main	Occurrence-bound discharge	7
M6	S08.main	Summary-grant scope	2
M7	S08.main	One-use grant consumption	3
M8	S10.main	Shared budget/quota	4
M9	S11.leak.bit0	Protected response release	1
M9	S11.leak.bit1	Protected response release	1
M10	S12.no_service	Evidence-backed remedy discharge	2

These mutants are deliberate omissions of known components and are classified as category 2 controls. They are not evidence that the unmutated B has a distinctive architecture.

The occurrence-replay control (M5) is the longest selected sensitivity witness at seven events: open; attest occurrence 1; discharge; reopen; present the old assertion; issue; dispatch. The old assertion fails to discharge the new occurrence; the corresponding positive control with fresh trusted evidence completes. This locates the tested distinction in ordinary occurrence-bound evidence enforcement, and B reproduced it.

Across the run, 16 selected minima were retained: 12 mutant-input witnesses and four unsupported-A full-profile witnesses. Reduction exhaustively considered the declared dependency-preserving subsequences in increasing event length with lexical tie-breaking. "Shortest" is therefore conditional on the original fixed fixture setup and reduction relation. The study does not claim a globally shortest counterexample or a minimum-cardinality setup.

4.6 Recording-pass limitation

The original runner retained all primary and mutant prefixes, the selected shortest reduction witnesses, and candidate counts. It did not retain a full transcript for every unsuccessful reduction candidate. After the outcome was known, a separately frozen recording pass replayed the same authorised reductions with unchanged experiment code, predicates, and first bad prefixes and wrote the missing candidate transcripts to reduction_recording_001/ALL_CANDIDATES.jsonl.

All 16 selected minima and their per-witness candidate counts matched the original reductions. The supplementary pass made 95 candidate replays, separate from the original 95 reduction replays, and those replays are excluded from primary and sensitivity-input denominators. This supports same-code reproducibility of the reduction record, but it is not independent replication, and the supplementary logs are not represented as if they had been captured during the original invocation.

5. Interpretation

5.1 The constructive null survives

The central result is that the strong conventional composition reproduced the tested r8 distinctions wherever actual r8 had supported coverage under the supplied semantics. No source-grounded irreducible enforceable r8 distinction survived the 57 comparable records or 308 comparable prefixes. Positive controls and mutant sensitivity make the result more informative than simple trace agreement because they show that the comparator was capable both of useful completion and of failing in the expected ways when named mechanisms were removed.

Within this experimental scope, the hypothesis that r8's tested consequential boundary requires a distinctive new architectural primitive is therefore not retained. The preregistered stopping rule applies. ROOT-0 remains NARROWED, r8 remains frozen and unchanged, r9-c1/c2 remain retired, and the distinctive-architecture escalation branch is stopped for this tested scope.

This is a substantive negative finding, not a failure of the investigation. Within this tested scope, the result blocks reasserting the same supported-projection distinctiveness claim merely under renamed terminology; a different claim would require new prospective evidence.

5.2 A useful integration profile can remain useful

Subsumption does not imply that the r8 integration profile has no engineering value. Bringing currentness, scoped grants, occurrence binding, stateful permissions, effect boundaries, observation, and remedy questions into one inspectable contract can still be useful for design, audit, training, or conformance testing. The result simply changes the defensible description: the present evidence supports an integration/conformance profile, not an irreducible mechanism.

A future claim of practical advantage would require a separate prospectively designed study of engineering cost, error rate, integration burden, latency, operator comprehension, portability, or another explicit utility measure. E1's local execution timings and finite synthetic corpus do not establish such an advantage.

5.3 The two gaps are coverage facts, not residual novelty

S11 and S12 are important because they demonstrate the discipline of preserving unsupported properties. If the paper had silently inserted a confidentiality filter or timed remedy worker into A, r8 would have been credited with properties absent from the frozen implementation. If the paper had instead counted the four diagnostics as ordinary A/B failures, it would have confused a known implementation coverage gap with evidence that a new architecture is required.

The correct interpretation is narrower. In E1, the finite S11 response-release relation and the stipulated S12 remedy workflow were instantiated in B using the declared conventional release-control and workflow/compensation mechanisms. Actual r8 v0.2-alpha.2-r8 does not implement those complete-profile properties. That fact prevents a full 12-family conformance claim but does not reopen the distinctive-architecture branch for the tested scope.

6. Threats to validity and limitations

Same-investigator construction. The investigator read r8 before designing B and the oracle. B is separately implemented but not independently conceived by a blinded team. The study is one AI-assisted investigation, not an independent replication.

Shared specification infrastructure. The fixture mapper, controllable plant, and reporting instruments are shared. A separate declarative oracle and fixed decision table reduce the risk that A and B merely agree with one another, but they do not eliminate common-mode specification error. The possibility that the supplied semantics omit an important real-world distinction remains open.

Synthetic authenticated inputs. E1 assumes authenticated domain records and trusted roots at the declared boundary. It does not test stolen credentials, compromised roots, hostile Python reflection, side channels, or arbitrary unobserved effects. Ordinary Python object identity in a demonstrator is not a deployment-grade capability system.

Controlled scheduling. The 272 scheduled post-setup events per arm are finite declared schedules. They are not an exhaustive operating-system race analysis or arbitrary interleaving theorem. Delayed notifications and uncontrolled service channels were excluded prospectively.

Reference composition gaps. Separate r8 formal slices and reference-code tests do not establish general composition or complete refinement. FORMAL-COMP-1, broader multi-effect and Lease integration, liveness composition, ontology/domain equivalence, and post-gate or exogenous-authority boundaries remain open.

Response confidentiality scope. S11's relation is demonstrated over a finite paired-input fixture and B's finite release rule. It is not a general theorem that B prevents all natural-language leakage or covert channels.

Remedy scope. S12 distinguishes partial external effect, restoration, independent witness, notice, and deadline. It does not prove that compensation can undo disclosure, social consequences, or arbitrary irreversible effects. Service availability remains a real assumption; the no-service fixture correctly ends overdue rather than claiming successful restoration.

Witness minimality. Selected reduction witnesses are shortest only under the fixed fixture setup and declared dependency-preserving subsequence relation. Setup cardinality was not minimised.

Post-outcome recording. The complete unsuccessful-candidate transcript archive was produced after outcome exposure. The pass used frozen code and reproduced the selected minima and counts, but it is not equivalent to having captured those transcripts prospectively and is not independent replication.

Unrun neighbours. Published neighbouring systems are literature context only. E1 did not pin, configure, or execute SARA, AADP, AAE, PCAA, Proof of Execution, Aegis, CPEX, CaMeL, or PACT. No failure score or superiority claim is assigned to them.

No institutional-legitimacy inference. Synthetic consent, purpose, authority, and remedy semantics are stipulations for the experiment. They do not establish legal power, ethical legitimacy, user consent, or jurisdiction-specific compliance in a deployment.

7. Reproducibility, evidence availability, and reporting discipline

The Stage 2 evidence package preserves the original handoff and Stage 1 package, the exact frozen E1 implementation and materialised corpus, pre-execution freeze, support matrix, original primary and mutant raw records, run identity, summary, verified accounting, unsupported-profile witnesses, mutant witnesses, and supplementary reduction-recording freeze/archive. The outer package manifest hashes every included member except itself. The original interpreter was Python 3.12.13 using the standard library.

This manuscript stage does not modify r8, the preregistration, experiment code, input corpus, oracle, raw result files, or witness records. It also does not rerun E1 to seek a different result. Any future reproduction must be labelled separately from the original run, including the environment and prior exposure.

The evidentiary hierarchy for PAPER-0 is: frozen/raw experiment records for empirical claims; frozen reference source and Stage 1 source-fidelity audit for implemented-r8 claims; primary publications for related-work claims; and the earlier QGIN-to-backtick history only as nonevidentiary provenance. Private specialist correspondence is not quoted or characterised here.

8. AI assistance, authorship, and correspondence

Meaningful AI assistance was used throughout ROOT-0/E1 for source inspection, protocol drafting, implementation support, analysis, accounting checks, and manuscript drafting. AI-assisted review was also used during manuscript review and publication preparation. The comparison was not a multi-investigator consensus simply because multiple model outputs or external specialists were consulted. Human responsibility for evidence interpretation and any eventual publication, authorship statement, source licences, and release decision remains with the project owner.

For this public manuscript, Weidu Fanling is the correspondence identity; no institutional affiliation is asserted. AI systems are not presented as authors, independent investigators, peer reviewers, or endorsers. No specialist outreach, private reply, or model agreement is presented as peer review or endorsement.

9. Conclusion

E1 v0.1 tested a strong constructive null for the frozen K0/CHI r8 consequential-action profile. Under identical supplied semantics and a controlled synthetic environment, a competent conventional composition matched actual r8 on every comparable input and prefix: 57/57 comparable records and 308/308 comparable prefixes. It also preserved useful completion, passing 26/26 positive-completion checks per arm, and the sensitivity suite detected all 12 designated mutant inputs for ten named mutants.

The full profile remains deliberately incomplete for actual r8. Two response-confidentiality diagnostics and two timed-remedy diagnostics preserve known implementation gaps, so no full 12-family conformance claim is warranted. Those gaps identify ordinary missing components rather than a surviving irreducible K0/CHI mechanism.

The supported conclusion is therefore bounded but decisive for the branch that E1 was designed to test: the tested r8 projection is subsumed by the declared conventional composition under supplied semantics, and the distinctive-architecture escalation branch stops for this scope. This does not prove universal equivalence, production safety, institutional legitimacy, or engineering superiority. For this tested scope, it prevents the same distinctive-architecture claim from being revived under new vocabulary without new prospective evidence.

References

1. J. H. Saltzer and M. D. Schroeder, "The Protection of Information in Computer Systems," Proceedings of the IEEE, 63(9), 1278-1308, September 1975. DOI: 10.1109/PROC.1975.9939. Author publication record: https://web.mit.edu/Saltzer/www/publications/pubs.html
2. A. W. Appel and E. W. Felten, "Proof-Carrying Authentication," Proceedings of the 6th ACM Conference on Computer and Communications Security (CCS 1999), 52-62, 1999. DOI: 10.1145/319709.319718. Author manuscript dated 9 August 1999: https://www.cs.princeton.edu/~appel/papers/says.pdf
3. J. Park and R. Sandhu, "The UCONABC Usage Control Model," ACM Transactions on Information and System Security, 7(1), 128-174, 1 February 2004. DOI: 10.1145/984334.984339.
4. OASIS, eXtensible Access Control Markup Language (XACML) Version 3.0, edited by Erik Rissanen, OASIS Standard, 22 January 2013. https://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html
5. A. Askarov and A. C. Myers, "Attacker Control and Impact for Confidentiality and Integrity," Logical Methods in Computer Science, 7(3:17), 1-33, 26 September 2011. DOI: 10.2168/LMCS-7(3:17)2011. https://lmcs.episciences.org/987
6. F. B. Schneider, "Enforceable Security Policies," ACM Transactions on Information and System Security, 3(1), 30-50, February 2000. DOI: 10.1145/353323.353382. Author copy: https://www.cs.cornell.edu/fbs/publications/EnfSecPols.pdf
7. R. Iannella and S. Villata, eds., ODRL Information Model 2.2, W3C Recommendation, 15 February 2018. https://www.w3.org/TR/2018/REC-odrl-model-20180215/
8. H. Garcia-Molina and K. Salem, "Sagas," Princeton CS-TR-070-87, January 1987; later SIGMOD 1987. Primary technical report: https://www.cs.princeton.edu/techreports/1987/070.pdf
9. C. K. Chow, "On Optimum Recognition Error and Reject Tradeoff," IEEE Transactions on Information Theory, 16(1), 41-46, 31 January 1970. DOI: 10.1109/TIT.1970.1054406
10. X. Guo et al., "When Tool Outputs Become Commands: Separating Action Induction from Runtime Authorization in Tool-Augmented LLM Agents," arXiv:2608.27146v1, submitted 27 August 2026 13:59:04 UTC. https://arxiv.org/abs/2608.27146v1
11. L. Fan et al., "The Granularity Mismatch in Agent Security: Argument-Level Provenance Solves Enforcement and Isolates the LLM Reasoning Bottleneck," arXiv:2605.11039v1, submitted 11 May 2026 04:09:11 UTC. DOI: 10.48550/arXiv.2605.11039. https://arxiv.org/abs/2605.11039v1
12. Y. Bengio et al., "Safety from Honesty in a Disinterested AI Predictor," arXiv:2606.29657v2, revised 10 July 2026 21:53:01 UTC; v1 submitted 28 June 2026 23:55:29 UTC. DOI: 10.48550/arXiv.2606.29657. https://arxiv.org/abs/2606.29657v2
13. E. Debenedetti et al., "Defeating Prompt Injections by Design," arXiv:2503.18813v2, revised 24 June 2025 08:05:33 UTC; v1 submitted 24 March 2025 15:54:10 UTC. DOI: 10.48550/arXiv.2503.18813. https://arxiv.org/abs/2503.18813v2
14. S. Saha, The Agent Action Decision Protocol (AADP): Per-Action Authorization for AI Agents, draft-saha-aadp-02, Internet-Draft, last updated 1 September 2026, work in progress. https://datatracker.ietf.org/doc/draft-saha-aadp/02/
15. L. K. Kroehl, Agent Authorization Envelope (AAE): A Machine-Evaluable Authorization Structure for Autonomous AI Agents, draft-kroehl-agentic-trust-aae-02, Internet-Draft, 6 September 2026, work in progress. https://datatracker.ietf.org/doc/draft-kroehl-agentic-trust-aae/02/
16. Z. Wang, "Proof-Carrying Agent Actions: Model-Agnostic Runtime Governance for Heterogeneous Agent Systems," arXiv:2606.04104v1, submitted 2 June 2026 18:10:35 UTC. The manuscript/source record also carries a 31 May 2026 heading; this paper uses the arXiv submission date for the version record. DOI: 10.48550/arXiv.2606.04104. https://arxiv.org/abs/2606.04104v1
17. J. Rhodes and G. Kang, "Proof of Execution: Runtime Verification for Governed AI Agent Actions," arXiv:2607.05397v1, submitted 26 April 2026 12:30:54 UTC. No chronology is inferred from the identifier prefix. DOI: 10.48550/arXiv.2607.05397. https://arxiv.org/abs/2607.05397v1
18. A. Mazzocchetti, "Runtime Governance for Agentic AI: Action-Boundary Control with Trusted Provenance and Fail-Closed Execution," arXiv:2608.16891v1, submitted 17 May 2026 23:44:12 UTC. This is the matching Aegis title; it is not the different March arXiv record considered during Stage 1 ambiguity resolution. DOI: 10.48550/arXiv.2608.16891. https://arxiv.org/abs/2608.16891v1
19. ContextForge, CPEX documentation, accessed 10 September 2026. Retrieved Rust documentation displayed version 0.2.2; no publication day is asserted. https://contextforge-org.github.io/cpex/ ; https://docs.rs/cpex/latest/cpex/ ; https://github.com/contextforge-org/cpex

Appendix A. Frozen empirical evidence map

Purpose	Frozen file / record
Prospective contract	inputs/ROOT0_E1_PAPER0_STAGE1_v0.1.zip -> E1_Complete_Contract_Preregistration_v0.1.md
Pre-execution code/input/oracle/support freeze	e1/PRE_EXECUTION_FREEZE.json
Outcome-free materialisation gate	e1/MATERIALISATION_CHECKS.json; e1/E1_MATERIALISATION_AUDIT_v0.1.md
Primary prefix evidence	e1/run_001/PRIMARY_RAW.jsonl
Mutant/sensitivity prefix evidence	e1/run_001/MUTANT_RAW.jsonl
Original run identity	e1/run_001/RUN_IDENTITY.json
Run summary	e1/run_001/SUMMARY.json
Read-only recount of recorded observations	e1/run_001/VERIFIED_ACCOUNTING.json
Unsupported full-profile diagnostics	e1/run_001/UNSUPPORTED_S11_*.json; e1/run_001/UNSUPPORTED_S12_*.json
Mutant minimal witnesses	e1/run_001/M*_WITNESS.json
Post-outcome reduction-recording declaration	e1/REDUCTION_RECORDING_FREEZE.json
Complete supplementary candidate transcripts	e1/reduction_recording_001/ALL_CANDIDATES.jsonl
Final bounded decision	E1_RESULT_AND_DECISION_RECORD_v0.1.md

Appendix B. Manuscript-stage integrity note

Stage 3 used the Stage 2 ROOT-0/E1 package and its nested Stage 1/original-handoff identities only. The broader QGIN archive was intentionally not used as technical evidence. The Stage 2 package identity is SHA-256 b320ed1f6e15764ff3ef7e8059624e1f2e897895c02a9820b6b7d18e87a226df for ROOT0_E1_STAGE2_RESULTS_v0.1.zip.

Stage 3 did not have direct filesystem access to the original Stage 2 raw byte stream and therefore preserved a fresh byte-level recount as a review blocker. During the Stage 4 cold manuscript review, the original Stage 2 ZIP bytes were directly available. The ZIP identity matched the pinned SHA-256, all 50 entries in MANIFEST.sha256 passed in a fresh extraction, and a fresh read-only recount of PRIMARY_RAW.jsonl, MUTANT_RAW.jsonl, SUMMARY.json, VERIFIED_ACCOUNTING.json, the four unsupported-profile records, the named witness files, and the reduction-recording archive confirmed 57 comparable inputs, 308 comparable prefixes, 26 positive-completion checks per arm, 12 designated sensitivity inputs across ten named mutants, and a separate post-outcome recording pass containing 95 candidate replays. Neither A nor B was executed during this recount.

The Stage 3 byte-access blocker is therefore resolved. This recount does not change the frozen E1 result, broaden its scope, or constitute independent-investigator replication.