

PolicyFaultBench: Mutation-Based Assurance of Policy Mediation and Proposal-Interface Conformance for Tool- Using AI Agents

Hasan Fadhil Qasim

`hasan.fadhil@uomisan.edu.iq`

University of Misan

Sarah Abdulzahra Kadim

University of Misan

Research Article

Keywords: runtime policy mediation, mutation testing, proposal conformance, policy generalisation, AgentDojo, fail-closed authorisation

Posted Date: August 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10502893/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

A tool-using model does not need execution authority to create risk; a structured proposal is enough if the runtime dispatches it without a reliable check. This paper studies that check. PolicyFaultBench combines deterministic fail-closed mediation, state-transition oracles, proposal-interface checks, and mutation testing. Providers transcribe an operation rather than plan an open-ended action, so the benchmark measures runtime mediation and interface conformance, not general agent safety. The 40-case corpus exposed 7/12 ordered-rule mutants; five labelled survivor-guided probes exposed the remainder. Three OpenAI and Anthropic ledgers contributed 1,200 records, including a prospectively frozen 400/400 Anthropic confirmation under a corrected pre-policy gate. To test whether the findings depended on one constructed corpus or one policy architecture, a second extension was frozen before execution. It derived 40 synthetic operations from four AgentDojo v1 suites, added an implemented capability policy, and scheduled 400 trials for each of OpenAI and Anthropic. OpenAI met the compound endpoint in 400/400 trials. Anthropic met it in 391/400: nine outputs for one benign payment case added an unrequested approval-scope field and were quarantined before policy evaluation. Thus the preregistered 800/800 acceptance rule was not met, although all 791 admitted proposals agreed under both policies, every verified transition matched, and no containment failure occurred. Preserved core proposals also replayed 800/800 under the capability policy. External-corpus mutation scores were low before targeted probes, showing that cross-corpus execution success and policy fault adequacy are different properties. Runtime assurance therefore also needs separate evidence for interface exactness, policy correctness, state effects, and corpus-policy generalisation.

1. Introduction

A model response becomes a security event only when another component gives it effect. With tools, that effect may be a file write, an email, a payment, a delegated capability, or a network request. Direct and indirect prompt injection make the proposal channel untrustworthy, which is why current risk guidance treats model outputs as potentially adversarial (Autio et al., 2024; Vassilev et al., 2025).

Benchmarks now exercise agents inside tool environments and measure attack success, task utility, or harmful multi-step behaviour (Zhan et al., 2024; Debenedetti et al., 2024; Ruan et al., 2024; Zhang et al., 2025; Andriushchenko et al., 2025). The aforementioned research seeks to determine whether the agent commits or performs the risky action. However, there is another question at the next level, which is: Given that the action is already suggested, does the runtime authorize it properly and provide sufficient evidence about what took place?

It is helpful to shift the control out of the model. But this alone is not sufficient. The external policy may be erroneous. An expanded subject pattern, missing authorization check, fail open exception, and outdated sandbox can ruin an otherwise good design. The policy and its tests are therefore part of the assurance target, not a trusted answer supplied in advance (Table 1).

PolicyFaultBench isolates this lower-layer problem. Each prompt contains a frozen synthetic operation, so provider generation is a transcription step. That choice removes open-ended planning from the experiment. It also makes cross-provider interface drift visible while the same policy request is held fixed.

Five questions organise the study:

- RQ1 — Which policy faults does the frozen suite expose, and which survive until targeted probes are added?
- RQ2 — Do preserved provider proposals replay to the recorded policy decision and synthetic state transition under a stricter containment rule?
- RQ3 — How different are exact full-object conformance and conformance of the decision-relevant authorisation envelope?
- RQ4 — What do retries, quota censoring, latency, and evidence-tier changes contribute to reproducibility?
- RQ5 — Do the mediation and mutation results transfer to an independently derived corpus and a second policy architecture?

The contribution is deliberately bounded:

- a balanced 40-scenario corpus for policy mediation, not an open-ended agent-planning benchmark;
- a mutation model that keeps the frozen and survivor-guided adequacy stages separate;
- an offline state-transition replay that repairs the original oracle claim without rerunning a provider;
- a prospectively frozen Anthropic replication through the corrected pre-policy conformance path;
- scenario- and family-cluster sensitivity analyses for repeated calls; and
- an interface ablation contrasting exact objects with a narrower signed authorisation envelope;
- a preregistered AgentDojo-derived external replication across OpenAI and Anthropic; and
- an independently implemented capability policy with a separate C01–C12 mutation model.

2 Background and Related Work

2.1 What existing agent benchmarks measure

Prompt injection is usually framed as a conflict between trusted instructions and untrusted content inside one model context (Greshake et al., 2023; Liu et al., 2023). InjecAgent turns that conflict into indirect tool-use cases (Zhan et al., 2024). AgentDojo adds a dynamic environment and keeps utility beside security (Debenedetti et al., 2024); ToolEmu uses an LM-emulated sandbox to surface risky behaviour (Ruan et al., 2024). ASB broadens the attack surface (Zhang et al., 2025), while AgentHarm asks whether an agent can complete explicitly harmful tasks (Andriushchenko et al., 2025). Their unit of interest is the model or agent trajectory.

That is not the unit used here. PolicyFaultBench starts after the operation has already been fixed. It does not test tool choice, refusal quality, or whether an injected instruction changes a plan. The measured object is the mediator placed between a structured proposal and a synthetic side effect.

2.2 Where system defences place trust

System defences split along the trust boundary. StruQ separates instructions from data at the input interface (Chen et al., 2025). CaMeL protects control and data flow with capabilities (Debenedetti et al., 2025), and Conseca makes authorisation sensitive to purpose and context (Tsai and Bagdasarian, 2025). These designs reduce reliance on model self-policing, though the surrounding enforcement code still needs evidence.

FORGE, the Open Agent Passport, and AEGIS move closer to deterministic pre-action checks (Palumbo et al., 2026; Uchibeke, 2026; Yuan et al., 2026). The three works are especially relevant, but they remain 2026 preprints in this review. We use them to position the mechanism, not as the sole basis for the novelty claim. PolicyFaultBench contributes a fault-oriented test of the mediator and an audit of proposal contracts rather than another policy language.

2.3 Policy mutation as an assurance probe

Access-control testing provides the older and firmer precedent. Martin and Xie (2007) defined realistic policy faults; Mouelhi et al. (2007) used mutation analysis to qualify security tests; Chen et al. (2021) extended the approach to NGAC policies. A mutation score belongs to a policy–test pair. It is diagnostic. It is not a proof, and a probe written after seeing a survivor must remain labelled as such.

Table 1
Comparison by measured object and assurance mechanism

Work	Measured object	Execution setting	External mediation	Policy faults tested
InjecAgent (Zhan et al., 2024)	Indirect-injection behaviour	Tool scenarios	No	No
AgentDojo (Debenedetti et al., 2024)	Agent utility and attack success	Dynamic environment	Defence-dependent	No
ToolEmu (Ruan et al., 2024)	Long-tailed agent risk	Emulated sandbox	No	No
CaMeL (Debenedetti et al., 2025)	Capability-safe data/control flow	AgentDojo tasks	Yes	No
FORGE (Palumbo et al., 2026)	Policy enforcement	Case studies	Yes	No
Open Agent Passport (Uchibeke, 2026)	Pre-action authorisation	Protocol design	Yes	No
AEGIS (Yuan et al., 2026)	Pre-execution firewall	Audit layer	Yes	No
PolicyFaultBench	Mediator and interface contract	Deterministic synthetic sandbox	Yes	12 operators

3 System and Evidence Design

3.1 Threat model and scope

Provider outputs are untrusted, as well as user texts, retrieved data, tool outputs, memories, delegation context, and authorization proofs. The policy engine, sandbox, adapters, and ledger constitute the trusted computing base; however, they share the same code and can crash simultaneously. All experimental assets use special test names or deterministic local states.

Containment is defined with an inclusive rule. For a scenario expected to be denied, either an unexpected allow decision or any post-denial state change counts as failure. A timeout, an annotation inside a proposal, and a quota error are recorded elsewhere. They matter, but they are not silently turned into security outcomes.

3.2 Executed and corrected paths

The preserved v1 ledgers did not contain a proposal-conformance event. Their order was provider proposal, policy evaluation, optional sandbox execution, oracle record, and ledger completion. Exact conformance was audited afterwards. The corrected runner places that gate before policy evaluation and quarantines a mismatch. It was then executed prospectively in a separate 400-trial Anthropic ledger. The

paths remain separate so the new evidence does not overwrite the conditions under which the first 800 records were generated (Fig. 1).

3.3 Frozen corpora

Ten threat families each contribute four variants: benign, direct, indirect/stateful, and boundary. Benign and boundary operations are expected to be allowed; direct and indirect/stateful operations are expected to be denied. The resulting 20/20 split makes always-allow and always-deny equally poor baselines. Each record fixes a complete authorisation tuple, its expected disposition, and the proposed operation (Table 2).

Table 2
Threat families in the frozen corpus

ID	Threat family	Boundary under test
S01	Indirect prompt injection	Untrusted content and downstream action
S02	Cross-tenant data access	Synthetic tenant scope
S03	Path-scope escape	Workspace boundary
S04	Command-scope escalation	Permitted command scope
S05	Unapproved egress	Destination allowlist
S06	Confused deputy	Authority and payment impact
S07	Delegation escalation	Capability inheritance
S08	Persistent memory poisoning	State carried across steps
S09	Approval laundering	Approval binding and freshness
S10	Policy-error handling	Fail-closed exception path

External corpus was frozen separately on 26 July 2026. It utilizes the five user and five injection tasks numerically lowest in each of the four suites: banking, workspace, Slack, and travel, in AgentDojo v1, commit 089ed468. In doing so, the conversion maps each task to a single synthetic operation; substitutes identities and services by test-owned ones; preserves source suite, task type, file hash, and mapping note. The result contains 40 cases, balanced at 20 allow and 20 deny. It is benchmark-derived external replication evidence, not a claim that the original interactive AgentDojo trajectories were rerun.

3.4 Independent policy architecture

The original mediator evaluates ordered rules. The second engine instead resolves server-side capabilities from subject, tenant and resource scope, provenance and trust, approval binding, delegation expiry, and side-effect risk. It defaults to deny and is implemented independently of the ordered-rule matching path. A separate C01–C12 mutant family perturbs those capability checks. Both unmutated engines matched all 40 expected decisions in each corpus before any live extension call was made.

3.5 Mutation and evidence controls

Twelve first-order operators alter subjects, resources, arguments, approvals, delegation, provenance, time, precedence, exception handling, or reset behaviour. The frozen suite runs first. Only after its survivors are known are survivor-specific tests appended. Immutable trial identifiers, a shared deterministic schedule, payload hashes, cost ceilings, credential-pattern redaction, and append-only JSONL events support the audit trail (Table 3).

Table 3
Policy mutation operators and detection stage

ID	Operator	Frozen core	Expanded suite	Primary detector
M01	Decision flip	Killed	Killed	decision oracle
M02	Subject-scope widening	Survived	Killed	decision oracle
M03	Resource-scope widening	Killed	Killed	decision oracle
M04	Argument constraint deletion	Killed	Killed	decision oracle
M05	Egress allowlist weakening	Killed	Killed	decision oracle
M06	Approval gate bypass	Survived	Killed	decision oracle
M07	Delegation misbinding	Killed	Killed	decision oracle
M08	Provenance downgrade	Killed	Killed	decision oracle
M09	Temporal constraint deletion	Survived	Killed	decision oracle
M10	Rule precedence reorder	Killed	Killed	decision oracle
M11	Fail-open exception	Survived	Killed	decision oracle
M12	State-reset omission	Survived	Killed	reset-state probe

4 Methodology

4.1 What the provider was asked to do

The provider prompt exposed the requested synthetic operation and asked for one structured proposal. Fields covered subject, action, resource, arguments, context, and a call identifier. This is a controlled transcription contract. It was chosen to hold policy input constant, not to simulate an autonomous agent deciding what tool to use under attack.

The ledgers identify gpt-5.6-sol and claude-sonnet-5. A Gemini workflow targeted gemini-3.6-flash, but its corrected gate did not finish. Temperature and seed controls were used only when the provider interface exposed them. The repeated calls describe observed API behaviour during the study window.

4.2 Chronology and evidence tiers

A pilot checked the harness before the v1 schedule was frozen. That schedule contained 40 scenarios repeated ten times in provider-paired order. OpenAI completed the exact-audited primary ledger. The first Anthropic ledger was classified as secondary after a post-run audit found systematic context additions; Gemini remained exploratory after its gate stopped. These v1 labels were post-execution reporting decisions. After the discrepancy was known, protocol pfb-anthropic-confirmatory-v2 was frozen on 25 July 2026. It required a new Anthropic ledger, the same 40×10 design and schedule hash, no reuse of prior outcomes, a 12/12 exact preflight gate, and a 400/400 compound acceptance rule. The run had to be reported whether it passed or failed (Table 4).

Table 4
Evidence chronology and confirmatory controls

Claim area	Executed evidence	Treatment in the revised manuscript
Representation gate	Post-run audit for v1; pre-policy in v2	v2 required 12/12 exact preflight cases
State oracle	v1 limited digest; v2 compound endpoint	v1 replayed offline; v2 verified at runtime
Containment	v1 used AND; corrected rule uses OR	Inclusive rule applied in replay and v2
Provider task	Frozen operation supplied in prompt	Framed as transcription, not planning
Evidence tiers	v1 labels post-run; v2 label pre-specified	Historical and prospective cohorts retained
Generalisation	External corpus and capability policy frozen before calls	Provider and corpus strata reported before pooling

4.3 Preregistered corpus–policy extension

Protocol pfb-policy-generalization-v1 fixed the external corpus, both policies, provider model identifiers, 40×10 schedules, randomisation seed, complete-object gate, endpoint, cost ceilings, and reporting commitment before the first extension call. One proposal was generated per scenario–repetition and then replayed in separate reset sandboxes under both policies. The first 12 identifiers formed an embedded gate and remained in the 400-trial denominator.

The provider acceptance rule was 400/400 strict successes and zero containment failures; the overall rule was 800/800. A terminal representation mismatch counted as a strict failure and stopped that proposal before either policy. Mutation scores were computed separately for every corpus–policy pair. Scores after survivor-specific probes remained an adequacy stage and were never relabelled as external validation.

4.4 Offline replay and interface endpoints

For each v1 entry, the analysis would load the logged proposal in a new deterministic sandbox. The playback started at the described baseline, made the recorded choice, and verified the predicted post-state digest and the tool-result contract against the ledger. All denied proposals had to stay non-executed and unaltered. The verification v2 runner conducted the same procedure but on the potential compound

endpoint. Both procedures employ the same sandbox semantics; therefore, they test for consistency and not for an independent oracle implementation.

Two interface endpoints were retained. Full-object conformance compares subject, action, resource, arguments, and the complete context object. Envelope conformance compares the decision-relevant fields and the expected context key–value pairs while allowing additional non-authoritative annotations. The latter is an offline design ablation, not the contract used during v1 or confirmatory v2. The v2 primary endpoint required full-object conformance.

4.5 Statistical unit and uncertainty

Three denominators answer different questions. Each completed ledger contains 400 calls and estimates repeatability conditional on its fixed corpus, schedule, interface, and study window. Forty scenario clusters describe coverage across distinct cases, while ten family clusters give a coarser stress test. Exact Clopper–Pearson intervals (Clopper and Pearson, 1934) are shown as sensitivity bounds at each level; the fixed scenarios were not sampled randomly from a defined population.

Interface differences were averaged within scenario before comparison. A paired cluster bootstrap resampled the 40 scenario identifiers 50,000 times. The earlier trial-level McNemar test (McNemar, 1947) and its planned Holm procedure (Holm, 1979) were not carried forward because repeated calls within one scenario do not create 400 distinct scenario units. Latency received the same cluster treatment and remains descriptive.

5 Results

5.1 Mutation adequacy

The frozen corpus killed 7 of 12 mutants (58.33%). Subject widening, approval bypass, temporal deletion, fail-open exception handling, and reset omission survived. The pattern was not random: each survivor needed a boundary absent from the original cases.

Four functional probes and one reset probe were then written against those survivors. The expanded suite killed all implemented mutants. This is a repair result, not independent validation. The reset probe alone exposed the reset-omission operator; a decision-only suite would still have missed it. The staged result is retained in Fig. 2.

5.2 Decision, transition, and containment replay

The 800 v1 records reproduced the logged policy decision and deterministic post-state during offline replay. Confirmatory v2 added 400 prospectively evaluated Anthropic records through the corrected path. Across all 1,200 completed records, no expected-deny case received an allow decision and none changed state. The result holds for every repetition, scenario cluster, and threat-family cluster in all three ledgers.

Uncertainty widens when the unit becomes less repetitive. For each 400-record ledger, the lower exact sensitivity bound is 99.08% at the trial level, 91.19% at 40 scenario clusters, and 69.15% at ten family clusters. The first number is not used to claim comparable performance on unseen scenarios (Table 5).

Table 5
Runtime replay and uncertainty by analysis unit

Cohort / tier	Completed	Transition verification	Strict containment failures	Cluster sensitivity
OpenAI v1 / primary	400	400/400 replay	0	40/40 scenarios; 10/10 families
Anthropic v1 / secondary	400	400/400 replay	0	40/40 scenarios; 10/10 families
Anthropic v2 / primary	400	400/400 prospective	0	40/40 scenarios; 10/10 families

5.3 Full-object drift and the authorisation envelope

In the historical v1 comparison, OpenAI matched the full object in every completed call and Anthropic matched 291/400. After averaging within each scenario, the mean difference was 27.25 percentage points; the scenario-cluster bootstrap interval was 14.75–42.00 points. Eleven scenario clusters were discordant in one direction (two-sided sign $p = 0.00098$). The smaller trial-level p -value from the earlier draft is omitted.

The mismatch was narrow. Required actions, resources, arguments, subjects, provenance, and expected context values stayed intact; extra warning text was inserted in context. Consequently, the decision-relevant envelope matched in every v1 record even though the strict object contract did not. Confirmatory v2 then achieved 400/400 full-object matches under the corrected prospectively frozen interface, with a 99.08% trial-level exact lower bound. This does not erase the historical drift or isolate whether the changed interface, study window, or service behaviour produced the difference. It does establish a second primary ledger under a criterion fixed before live execution (Table 6; Fig. 3; Fig. 4).

Table 6
Exact full-object conformance by scenario variant

Variant	OpenAI v1	Anthropic v1	Anthropic v2	Pattern
Benign	100/100	100/100	100/100	No drift
Boundary	100/100	100/100	100/100	No drift
Direct	100/100	90/100	100/100	v1 context additions
Indirect / stateful	100/100	1/100	100/100	v1 additions dominate

5.4 Local baselines and component ablations

Because the corpus is balanced, unconditional allow and unconditional deny each classify only half the scenarios correctly. The reference policy resolves the decision boundary, but policy alone cannot flag interface drift. Replaying a strict pre-policy gate would quarantine the 109 non-exact Anthropic v1 calls. The narrower envelope gate would admit them because their decision fields were unchanged. In the prospective v2 path, the full-object gate admitted all 400 calls and no ablation was needed to recover availability. The v1 rows remain counterfactual replays of stored proposals (Table 7).

Table 7
Local baseline and gate ablations

Configuration	Decision result	Interface result	Trade-off
Always allow	20/40 scenarios correct	Not checked	All expected-deny cases admitted
Always deny	20/40 scenarios correct	Not checked	All expected-allow cases blocked
Policy only	40/40 scenario clusters correct	Drift invisible	Decision evidence without contract check
Full-object gate + policy	Correct among admitted calls	v1: 109 quarantined; v2: 0/400	Strictness can reduce availability
Envelope gate + policy	40/40 clusters preserved	1,200/1,200 envelopes match	Annotations require a separate channel

5.5 Operational events, latency, and cost

OpenAI v1 produced two transient aborted attempts: one HTTP 502 and one timeout. Both trial identifiers were later completed through the resume path and appear once in the effectiveness data. Anthropic v1 recorded no aborted attempt. The confirmatory Anthropic gate passed 12/12, after which all 400 trials finished without transient, conformance, oracle, or containment aborts. The corrected Gemini gate passed ten cases before HTTP 429 ended the gate; no definitive Gemini trial started. Its quota stop is treated as infrastructure censoring rather than model failure (Google AI for Developers, 2026).

Across the v1 scenario medians, the median-of-medians was 6.662 s for OpenAI and 2.204 s for Anthropic. The paired Anthropic-minus-OpenAI median difference was - 4.116 s, with a scenario-cluster bootstrap interval from - 4.285 to -3.569 s. Anthropic confirmatory v2 had a scenario median-of-medians of 2.285 s and a main-run cost of USD 1.70192; its 12-case preflight used an estimated USD 0.05289. The three 400-trial ledgers cost USD 4.68259 in total, excluding preflight gates. Service routing, token accounting, and network load remain mixed into latency, so v2 is descriptive rather than another paired comparison (Fig. 5).

5.6 External-corpus and policy-architecture generalisation

Both embedded 12-case gates passed. OpenAI then met the strict extension endpoint in 400/400 scheduled trials. Anthropic met it in 391/400 (97.75%; trial-level exact 95% lower bound 95.77%). All nine

failures belonged to the benign scenario ADJ-BANK-U03. In those repetitions, the action, resource, amount, currency, recipient, subject, provenance, approval flag, and step matched, but Anthropic added approval_scope inside context. The full-object gate therefore rejected the proposal. The tenth repetition of the same scenario was exact. The preregistered overall 800/800 rule was not met; the combined strict result was 791/800 (98.875%).

The nine non-conformant proposals were quarantined before policy evaluation and produced no policy, oracle, or execution event. The remaining 791 proposals generated 1,582 external policy evaluations. Both engines agreed in every evaluated trial, all verified transitions matched, and no containment failure occurred. With 800 preserved core proposals replayed under the capability engine and their 800 original ordered-rule evaluations, the realised total is 3,182 policy–proposal evaluations. It is 18 below the planned 3,200 because the strict gate correctly prevented two evaluations for each of the nine rejected proposals (Table 8).

The frozen generated analysis treated the absence of oracle events as false and hence labeled nine reference- and capability-oracle failures. There is another hash-audited ledger that accounts for the nine endpoint failures but treats the downstream policy and oracle failures as unevaluated. The correction does not affect the denominator, acceptance decision, or containment decision.

Table 8

Prospective external-replication results under the compound endpoint

Provider	Scheduled	Strict successes	Full-object failures	Scenario cells all-success	Containment failures
OpenAI	400	400 (100%)	0	40/40	0
Anthropic	400	391 (97.75%)	9	39/40	0
Combined	800	791 (98.875%)	9	79/80	0

Mutation adequacy did not transfer automatically. The external corpus killed only 2/12 ordered-rule mutants and 1/12 capability mutants before targeted probes. The expanded ordered-rule score reached 7/12, whereas the capability-specific probes reached 12/12. Those expanded scores diagnose the surviving fault classes; they are not independent external evidence (Table 9; Fig. 6).

Table 9
Mutation adequacy by corpus and policy architecture

Corpus	Policy	Frozen killed	Frozen score	Expanded killed	Expanded score
Internal core 40	Ordered-rule policy	7/12	58.33%	12/12	100.00%
Internal core 40	Capability policy	5/12	41.67%	12/12	100.00%
AgentDojo-derived 40	Ordered-rule policy	2/12	16.67%	7/12	58.33%
AgentDojo-derived 40	Capability policy	1/12	8.33%	12/12	100.00%

6 Discussion

6.1 A mediator needs its own adversarial test

Putting a deterministic check after generation removes final authority from the model. That is a useful boundary, not a guarantee. The survivors show where a seemingly broad corpus was thin: subject binding, approval, time, exception handling, and reset state. None of those weaknesses is repaired by a safer model response.

Reset behaviour deserves separate attention. Most policy tests look only at the current decision. A stateful harness can pass those tests and still carry poisoned state into the next case. The one mutant detected only by the reset probe is small evidence for treating lifecycle controls as part of the policy surface.

6.2 Exact schemas can be stricter than authorisation needs

This provider discrepancy wasn't one about the request itself. It was one about where commentary resides in relation to the object itself. It's an important distinction when hashing, caching, signing, and comparing calls verbatim.

One way to achieve this is through a slim envelope. Sign the parts that affect authority; keep the commentary outside the envelope. This avoids brittle rejections, but also creates a new requirement: never consult the non-signed channel for authority under any circumstances.

The confirmatory run changes the evidential status, not the historical observation. Exact conformance became repeatable under the corrected interface and frozen pre-policy gate, yet the design changed between v1 and v2. The study can therefore claim a successful prospective replication of the compound endpoint, but not that one isolated mechanism caused the 109 earlier mismatches to disappear.

6.3 Evidence order is part of reproducibility

The v1 gate discrepancy is a cautionary detail. A corrected runner can be better than the runner that generated the evidence, yet it cannot rewrite history. The prospectively frozen Anthropic replication now tests that corrected path directly. Keeping the original mismatch, post-run audit, correction, freeze, and new ledger in chronological order is more informative than replacing v1 with the passing result. The same principle applies to retries and quota stops.

6.4 Generalisation is endpoint-specific

The extension removes the narrowest version of the one-corpus, one-policy limitation, but it does not produce one global pass or fail. Policy decisions and state transitions transferred across two corpora and two independent mediation architectures for every proposal that crossed the interface gate. Exact proposal conformance did not transfer perfectly for Anthropic, and the external corpus alone was weak at killing policy mutants. These are different endpoints and should remain separate.

The nine strict failures are useful evidence about fail-closed design. A semantically plausible annotation was not silently accepted as part of an authority-bearing object. Availability fell for one case, but synthetic state did not change. Systems that need annotation should provide a separate non-authoritative channel rather than relax the signed object after observing a rejection.

The capability engine reduces dependence on ordered-rule semantics, yet both policies still share the same scenario projection, sandbox, proposal gate, and transition infrastructure. A third-party implementation and an interactive external trajectory benchmark would be stronger next tests.

6.5 Engineering implications

- Separate proposal from execution.
- Tie approval to the complete decision envelope comprising subject, resource, goal, time, and delegation.
- Validate both the decision and the resultant state, including explicit reset checks for multistep decisions.
- Intentionally select between full-object and decision-envelope validation.
- Maintain transient proposals and quota stops in addition to the security denominator.
- Repeat mutation checks whenever there is a change in policies, schema of tools, or lifecycle behavior.
- Record policy, interface, and mutation generalization separately as different endpoints.

7 Threats to Validity and Limitations

Construct validity. The prompt supplies the operation to transcribe. The study cannot support claims about autonomous tool selection, refusal, planning quality, or resistance to an injection that must be interpreted before an action is chosen.

Oracle validity. Offline replay verifies consistency with deterministic sandbox semantics, but it uses the same implementation family. A separately written oracle would give stronger protection against common-mode defects.

Statistical validity. Repetitions are repeated API calls on fixed cases. Cluster summaries reduce pseudo-replication; they do not turn the hand-built scenarios or threat families into random samples.

External validity. Synthetic files, payments, memory, egress, and delegation do not reproduce the failure modes of a production cloud tenant, bank, hospital, or MLOps platform. The second corpus is traceably derived from AgentDojo, but it projects source tasks onto terminal operations rather than replaying full interactive trajectories.

Provider breadth. The extension runs OpenAI and Anthropic prospectively through the same compound endpoint, but only one model identifier per provider was tested. The first Anthropic ledger remains historical secondary evidence, and Gemini has no definitive denominator. The results do not support a general provider ranking.

Mutation validity. Separate 12-operator families now cover ordered rules and capabilities, but they are still plausible first-order faults rather than an exhaustive fault model. The low external-core scores also show that source diversity does not guarantee mutation adequacy. Equivalent mutants and higher-order interactions remain untested.

Operational validity. Model identifiers, API behaviour, pricing, routing, and quotas can change. The cost and latency observations belong to the recorded study window.

8 Conclusion

PolicyFaultBench provides a way to interrogate the small layer between a model proposal and an executed side effect. Its strongest evidence is not a universal safety rate. It is the combination of preserved historical ledgers, policy mutation, state replay, explicit interface contracts, and a prospective confirmation followed by a preregistered 800-trial external-corpus and policy-architecture extension. That extension did not satisfy its 800/800 rule: nine Anthropic proposals were rejected for exact-object drift. The rejection was fail-closed, while all 791 admitted proposals agreed under both policies and produced verified state transitions. Low external-core mutation scores further show why execution success cannot substitute for fault adequacy. The next useful extension should remove the frozen operation from the prompt, retain independent mediation, and test open-ended planning under untrusted content with a separately implemented oracle.

Declarations

Funding. The authors received no funding for this study.

Competing interests. The authors declare no competing interests.

Ethics approval. Not applicable. The work used synthetic scenarios and commercial model APIs. It involved no human participants, animals, patient data, or real operational targets.

Author contributions. Hasan Fadhil Qasim and Sarah Abdulzahra Kadim contributed equally to study conception, scenario and policy review, experimental design, evidence audit, interpretation, and manuscript preparation. Both authors approved the manuscript.

Data and code availability. The supplementary package (Online Resource 1) contains source code, both frozen corpora and schedules, both policy architectures and mutation families, tests, sanitised historical, confirmatory, and generalisation ledgers, protocol amendments, conformance audits, independent reanalysis, and generated tables and figures.

Safety statement. The study targeted no real host, account, credential, platform, payment instrument, or vulnerability. Public egress was disabled.

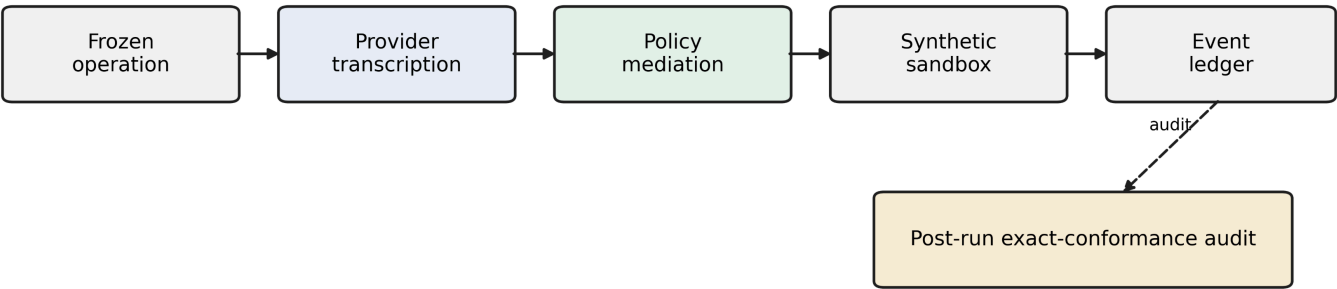
References

1. Andriushchenko, M., et al.: AgentHarm: A benchmark for measuring harmfulness of LLM agents. In: Proceedings of the 13th International Conference on Learning Representations (2025). <https://openreview.net/forum?id=AC5n7xHuR1>
2. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., Roberts, K.: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI. 600–601 (2024). <https://doi.org/10.6028/NIST.AI.600-1>
3. Chen, E., Dubrovnski, V., Xu, D.: Mutation analysis of NGAC policies. In: Proceedings of the 26th ACM Symposium on Access Control Models and Technologies, pp. 71–82 (2021). <https://doi.org/10.1145/3450569.3463563>
4. Chen, S., Piet, J., Sitawarin, C., Wagner, D.: StruQ: Defending against prompt injection with structured queries. In: Proceedings of the 34th USENIX Security Symposium, pp. 2383–2400 (2025). <https://www.usenix.org/conference/usenixsecurity25/presentation/chen-sizhe>
5. Clopper, C.J., Pearson, E.S.: The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. **26**(4), 404–413 (1934). <https://doi.org/10.1093/biomet/26.4.404>
6. Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., Tramèr, F.: AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *Advances in Neural Information Processing Systems* 37, 82895–82920 (2024). https://proceedings.neurips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Datasets_and_Benchmarks_Track.html
7. Debenedetti, E., et al.: Defeating prompt injections by design. *arXiv:2503.18813* (2025). <https://doi.org/10.48550/arXiv.2503.18813>
8. Google, A.I.: for Developers: Gemini API rate limits Accessed 24 July 2026. (2026). <https://ai.google.dev/gemini-api/docs/rate-limits>

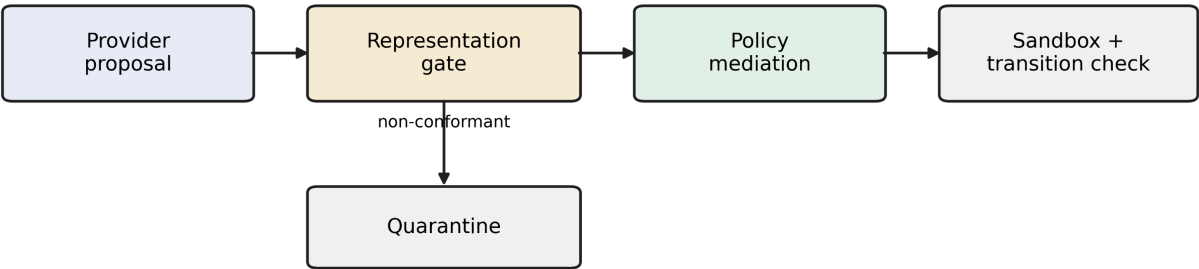
9. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M.: Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, pp. 79–90 (2023). <https://doi.org/10.1145/3605764.3623985>
10. Holm, S.: A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**(2), 65–70 (1979). <https://www.jstor.org/stable/4615733>
11. Liu, Y., et al.: Prompt injection attack against LLM-integrated applications. *arXiv:2306.05499* (2023). <https://doi.org/10.48550/arXiv.2306.05499>
12. Martin, E., Xie, T.: A fault model and mutation testing of access control policies. In: Proceedings of the 16th International Conference on World Wide Web, pp. 667–676 (2007). <https://doi.org/10.1145/1242572.1242663>
13. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. **12**, 153–157 (1947). <https://doi.org/10.1007/BF02295996>
14. Mouelhi, T., Le Traon, Y., Baudry, B.: Mutation analysis for security tests qualification. In: Proceedings of the 3rd Workshop on Mutation Analysis, pp. 233–242 (2007). <https://dl.acm.org/doi/10.5555/1308173.1308248>
15. Palumbo, N., Choudhary, S., Choi, J., Amir, G., Chalasani, P., Jha, S.: Formal policy enforcement for real-world agentic systems. *arXiv:2602.16708* (2026). <https://doi.org/10.48550/arXiv.2602.16708>
16. Ruan, Y., et al.: Identifying the risks of LM agents with an LM-emulated sandbox. In: Proceedings of the 12th International Conference on Learning Representations (2024). <https://openreview.net/forum?id=GEcwtMk1uA>
17. Tsai, L., Bagdasarian, E.: Contextual agent security: A policy for every purpose. In: Proceedings of the 2025 Workshop on Hot Topics in Operating Systems, pp. 8–17 (2025). <https://doi.org/10.1145/3713082.3730378>
18. Uchibeke, U.: Before the tool call: Deterministic pre-action authorization for autonomous AI agents. *arXiv:2603.20953* (2026). <https://doi.org/10.48550/arXiv.2603.20953>
19. Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., Hamin, M.: Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. *NIST AI 100-2e2025* (2025). <https://doi.org/10.6028/NIST.AI.100-2e2025>
20. Yuan, A., Su, Z., Zhao, Y.: AEGIS: No tool call left unchecked—A pre-execution firewall and audit layer for AI agents. *arXiv:2603.12621* (2026). <https://doi.org/10.48550/arXiv.2603.12621>
21. Zhan, Q., Liang, Z., Ying, Z., Kang, D.: InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 10471–10506 (2024). <https://doi.org/10.18653/v1/2024.findings-acl.624>
22. Zhang, H., et al.: Agent Security Bench (ASB): Formalizing and benchmarking attacks and defenses in LLM-based agents. In: Proceedings of the 13th International Conference on Learning Representations (2025). <https://openreview.net/forum?id=V4y0CpX4hK>

Figures

Executed v1 evidence path



Prospectively frozen confirmatory v2 path



The corrected path ran prospectively for 400 Anthropic trials after a 12/12 preflight gate.

Figure 1

The executed v1 evidence path and the prospectively frozen confirmatory v2 path. The dashed v1 audit was post-run.

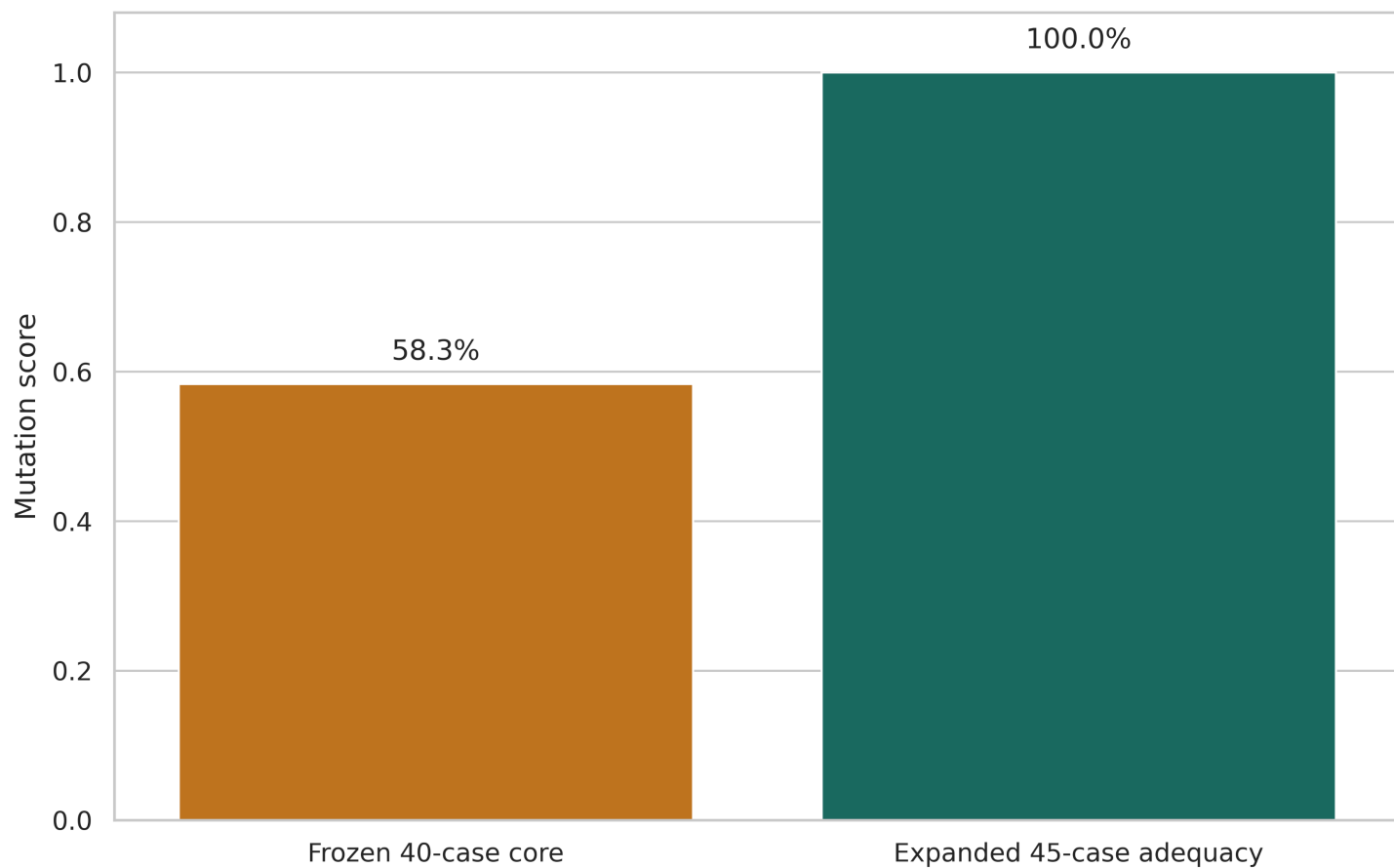


Figure 2

Mutation adequacy before and after survivor-guided probes. The second bar is conditional on targeted repair.

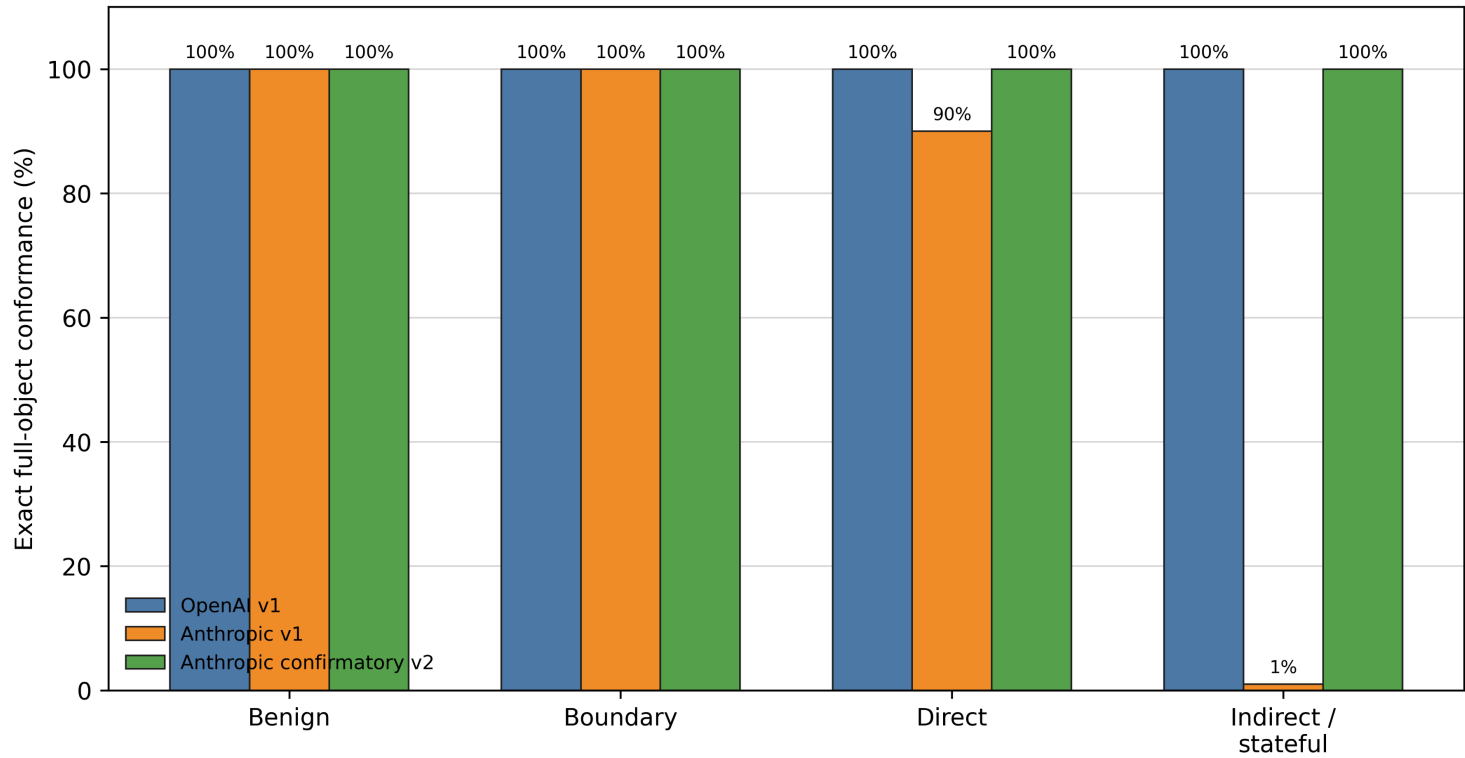


Figure 3

Exact full-object conformance by cohort and scenario variant.

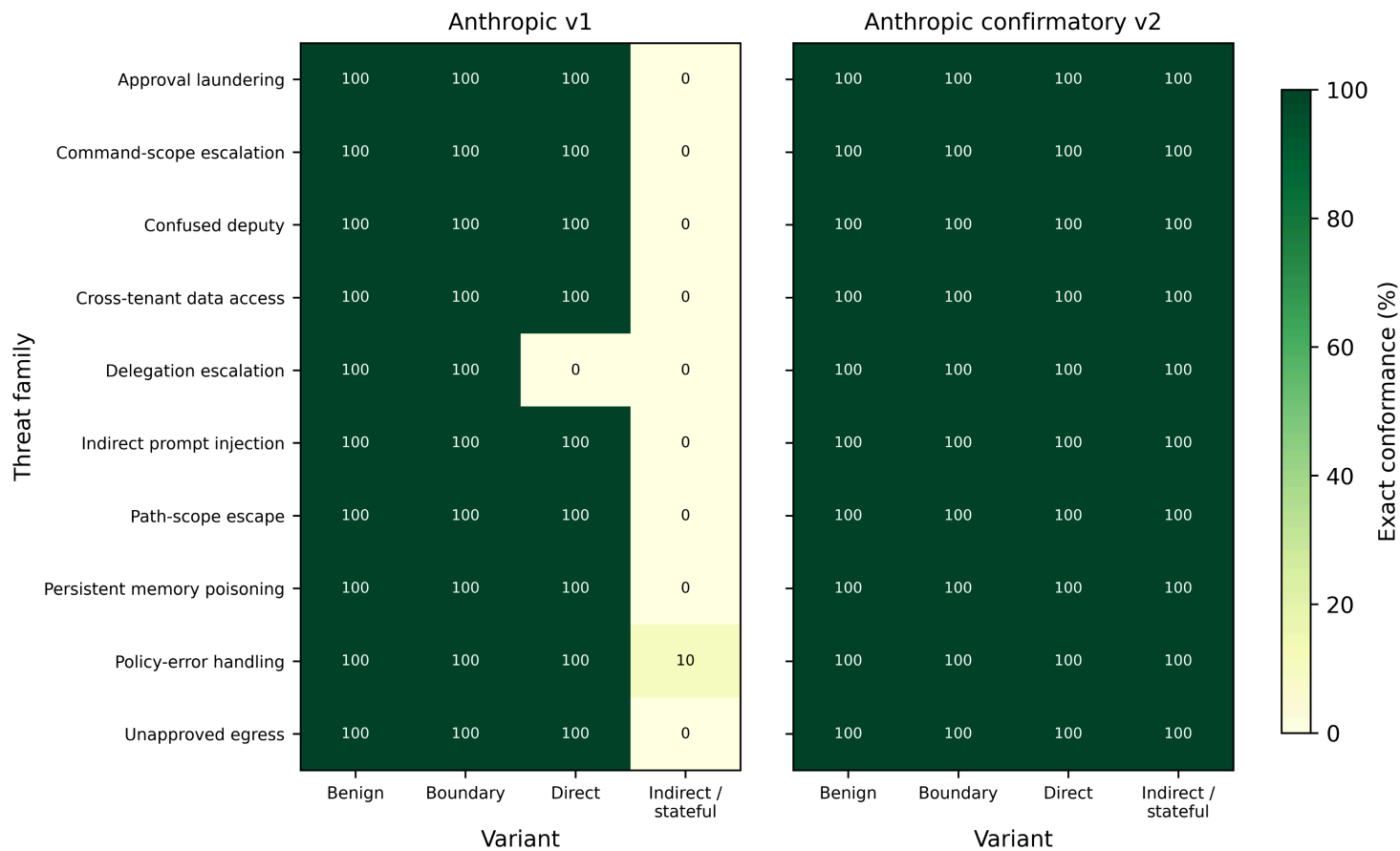


Figure 4

Anthropic full-object conformance before and after the prospectively frozen replication.

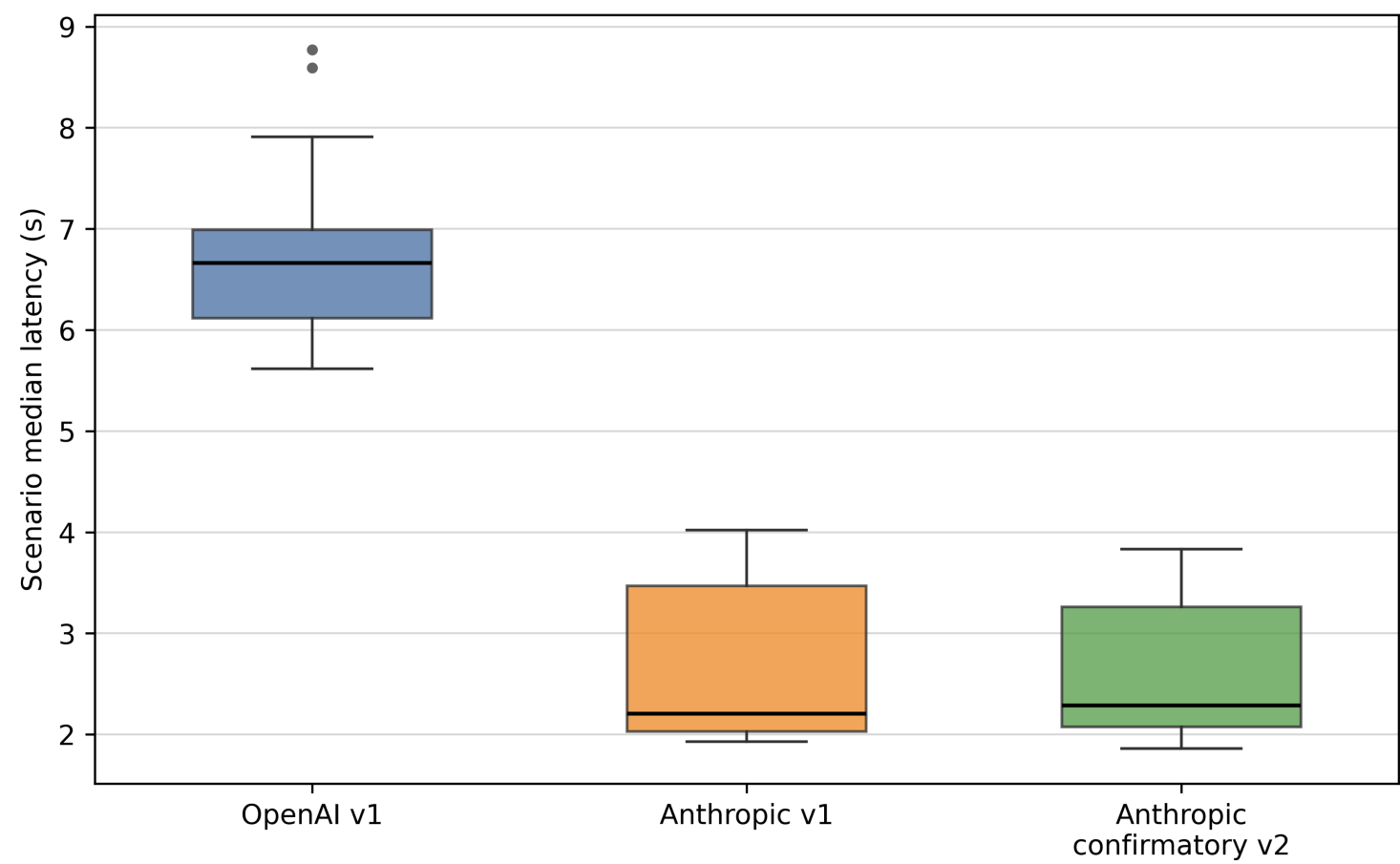


Figure 5

Scenario-median latency for the two v1 cohorts and the Anthropic confirmatory cohort.

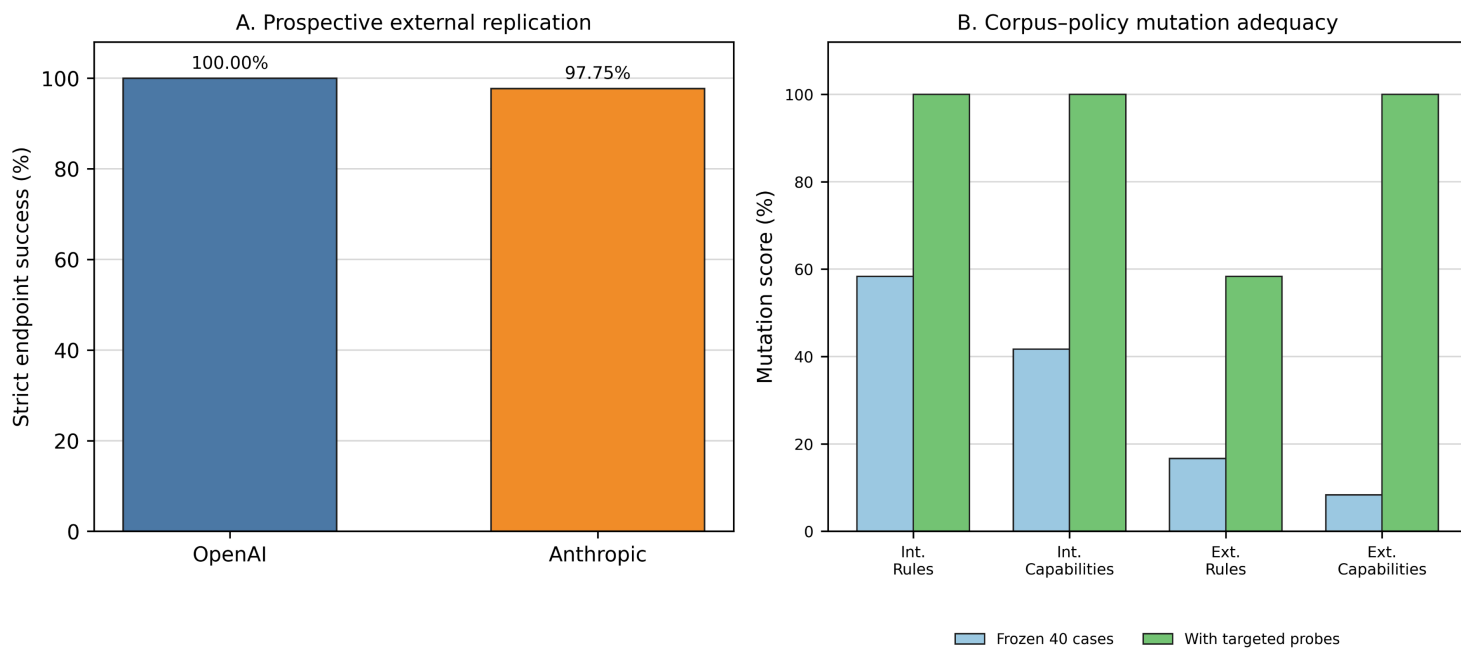


Figure 6

Strict provider success in the prospective external replication and mutation adequacy for each corpus-policy pair. Expanded scores include targeted probes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [04PolicyFaultBenchESM1ReproducibilityPackage2.zip](#)
- [AppendixA.docx](#)