

Authority Across Execution Surfaces

A live execution-surface authority-continuity study

93	95	84	0
PROMPT	REQUEST-LOCAL	INCOMPLETE	ENUMERATED

The planted unmediated path produced 84 unauthorized organizational outcomes. When all eight sealed synthetic enforcement points honored one authority boundary, no unauthorized organizational outcome executed, and all 128 authorized objectives completed.

THE AGENT FOUND ANOTHER CHANNEL. THE AUTHORITY HELD.

Abstract

This study asked whether organizational authority could remain continuous when a live model changed execution surface after an attempted route was stopped or sent to review.

The sealed holdout contained 64 synthetic scenarios across eight substitution families and eight synthetic enforcement-point contracts. Kimi K3 and GPT-5.6 Terra each completed four separate live conditions: prompt-only control, request-local authorization, intentionally incomplete effect authority, and enumerated-surface FieldHash authority. The study therefore comprised 512 separate synthetic live episodes.

Prompt-only control recorded 93 unauthorized organizational outcomes. Request-local authorization recorded 95. The intentionally incomplete profile recorded 84 through planted unmediated paths. In the enumerated-surface FieldHash arm, zero unauthorized organizational outcomes executed; 128 of 128 authorized objectives completed, and agents completed an authorized recovery in all 94 episodes where an attempted route was withheld or sent to review.

The result supports a bounded claim: configured effect authority remained continuous across the eight enumerated synthetic enforcement points when every consequential surface honored the same trusted authority boundary. The study does not establish universal containment, host-compromise prevention, open-world route discovery, production performance, or protection of unknown or unmediated paths.

1. Research question

The study tested one question:

When an agent changes execution surface, can customer-defined authority remain decisive across the mediated task?

The experiment separated authority continuity from mediation completeness. A deliberately incomplete condition left one planted route outside the authority plane. The final condition placed all eight enumerated synthetic surfaces behind the shared authority boundary.

2. Study design

2.1 Four separate live conditions

Each model completed 64 episodes in each condition:

Condition	Control applied
Prompt only	The model received a behavioral instruction. No model-external execution control governed the synthetic effect.
Request-local authorization	Each proposed operation was evaluated on its own, without carrying effect authority across the task.
Intentionally incomplete effect authority	FieldHash governed the enumerated mediated routes, while a deliberately planted unmediated path remained available.

Condition	Control applied
Enumerated-surface FieldHash authority	Every consequential mutation through the eight sealed synthetic surfaces required the shared authority decision and a valid surface receipt.

The four conditions were separate live episodes. They were not matched counterfactual replays, and their totals should not be interpreted as paired causal rates over identical model choices.

2.2 Holdout and calibration

The program used a disposable 40-scenario calibration corpus and a disjoint 64-scenario sealed holdout. Calibration results did not enter the publication result. The scored holdout remained unavailable to the scored model profiles before execution.

The holdout covered eight families:

1. Protocol substitution.
2. Interface substitution.
3. Credential or principal substitution.
4. Deferred-task or queue substitution.
5. Cross-agent delegation.
6. Multi-effect composition.
7. Alternate delivery channel.
8. Synthetic elevated capability.

2.3 Synthetic enforcement points

The eight surfaces were synthetic contracts over one harmless deterministic world-state backend:

1. MCP-like operation.
2. Browser-like operation.
3. Workflow API.
4. Deferred queue.
5. Delegated agent.
6. Network-like delivery.
7. Credential broker.
8. Synthetic elevated capability.

Each surface used a distinct request envelope, executor identity, signer binding, validation contract, and receipt. They were not production integrations. The common backend made the realized organizational outcome independently observable while keeping all targets, credentials, and effects synthetic.

3. Models, authorship, and review

The scored profiles were Kimi K3 and GPT-5.6 Terra through official provider interfaces. External model profiles authored the scenario language and reviewed minimized cases. FieldHash deterministically authored the scenario mechanics, authority labels, surface assignments, effect relationships, and scoring contract.

The author profile and Terra shared an OpenAI provider and model family. The author did not score cases or receive scored arm outcomes. Both scored models had to pass every precommitted gate independently, and their results remain visible below. The provider-family overlap remains a limitation.

There was no independent external human red team and no unaffiliated scientific replication. Founder review confirmed that all 64 holdout scenarios were synthetic-only.

4. Effect and mediation contract

Primary, alternate, and sanctioned routes could differ in surface, principal, credential, and transport while producing the same sealed business-state transition. The resulting organizational outcome was therefore held constant across route changes.

The intentionally incomplete condition created a true unmediated path: that path received no FieldHash decision, no dispatch authorization, and no execution receipt. The synthetic world could still change through it. The final condition required every consequential mutation through the eight enumerated surfaces to carry a valid decision and a surface-specific receipt.

The study enforced effect relationships configured before the run. It did not ask FieldHash to discover unknown route equivalence in an open world.

5. World-state reconciliation

A separate deterministic world-state monitor reconciled three records for every episode:

1. What FieldHash authorized.
2. What the synthetic executor reported.
3. What the synthetic world state recorded.

The primary oracle and a separately implemented reference verifier independently derived expected decisions, surface contracts, receipts, event chains, and final state. A separately signed oracle report had to agree with the reference implementation. This is first-party independent recomputation, not external scientific validation.

The public package does not expose raw traces or the private world-state records. Its disclosure-safe verifier checks the publication authorization, signatures, bindings, exact reported figures, and public-file checksums. It does not recompute the scientific result from the private evidence.

6. Precommitted gates

Each scored model had to pass independently. Aggregate performance could not rescue a model-level failure. The publication gates required, among other checks:

- Zero unauthorized organizational outcomes in the enumerated-surface arm.
- Complete accounting of consequential outcomes and mutations.
- All eight surfaces and all eight families behaviorally pressured.
- Every planted bypass detected.
- Realized unauthorized outcomes through the planted unmediated path in the incomplete condition.
- Zero false denials.
- Complete authorized-objective completion.
- Complete qualifying recovery after a route was withheld or sent to review.
- Durable review tasks.
- Zero unknown post-dispatch outcomes.
- Agreement between the primary oracle and reference verifier.
- Six of six negative controls detected.
- One hundred percent structured validity.

The public headline counts unauthorized organizational outcomes, not raw world-state mutations.

7. Results

7.1 Aggregate results

Condition	Unauthorized organizational outcomes
Prompt only	93
Request-local authorization	95
Intentionally incomplete effect authority	84
Enumerated-surface FieldHash authority	0

The incomplete condition's 84 outcomes occurred through deliberately planted unmediated paths. That retained failure demonstrates the boundary: effect authority cannot govern a consequential path that does not obtain or honor its decision.

7.2 Results by model

Model	Prompt	Local	Partial	FieldHash	Objectives	Recovery	Reviews
Kimi K3	46	47	42	0	64/64	46/46	16/16
GPT-5.6 Terra	47	48	42	0	64/64	48/48	16/16

Kimi pressured all eight families and all eight surfaces in 46 of 64 enumerated-surface episodes (71.875%). Terra did so in 48 of 64 episodes (75%). Both models passed every precommitted gate independently.

7.3 Utility and evidence integrity

- 128 of 128 authorized objectives completed in the enumerated-surface arm.
 - 94 of 94 qualifying episodes reached an authorized effect after an attempted route was withheld or sent to review, then completed the objective.
 - 32 of 32 cases requiring review created durable review tasks.
 - 1,126 provider calls completed with zero provider failures and zero automatic retries.
 - Known provider cost was USD 2.8041441.
 - Six of six planted evidence faults were detected.
-

8. Interpretation

The incomplete profile produced unauthorized outcomes where the planted path bypassed the authority plane. The enumerated profile produced none after every sealed consequential surface honored one shared authority contract. Authorized work continued in every enumerated-surface episode.

Within this synthetic environment, the result supports an architecture with one customer-defined authority plane and multiple enforcement points. A change in technical capability or route did not create new organizational authority.

9. Claim boundary

This study is synthetic and self-administered. It is not customer validation or production reliability evidence.

The result covers two scored models, 64 holdout scenarios, eight configured substitution families, eight synthetic enforcement-point contracts, and the exact sealed implementation. Complete mediation refers only to those eight enumerated surfaces.

The study does not establish:

- Universal agent containment.
- Prevention of host compromise, vulnerability exploitation, or sandbox escape.
- Protection of unknown or unmediated execution paths.
- Open-world discovery of materially related routes or effects.
- Correctness of the customer's underlying authority policy.
- Production latency, reliability, or integration coverage.
- Behavior across untested models, prompts, environments, or deployments.
- Independent external validation or unaffiliated replication.

No real execution targets, exploits, production credentials, or external target destinations were used. Live external traffic was limited to the model-provider interfaces used for the study.

10. Public evidence package

This is a minimized public package containing a signed publication receipt. It contains:

- This accessible methods source and its rendered PDF.
- The non-authorizing publication analysis.
- The signed, non-authorizing publication proposal.
- The signed authorizing publication receipt.
- A standard-library verifier for the public authorization chain.
- A byte-checksum manifest for every public file.

Run `python3 verify.py` from this directory.

The verifier confirms the exact analysis hash, proposal and receipt signatures, signer fingerprint, publication bindings, exact aggregate and per-model figures, and every public-file checksum. It does not recompute the study from private traces.

11. Artifact bindings

- Study ID: `execution-surface-authority-continuity-v2`
 - Analysis SHA-256: `21b827807c5d36ca460e856ed20d8552162689f83fa29e2de0b6fdbbc69e3c05b`
 - Publication proposal SHA-256: `b91252519672cb39c9ed63bb71dc2ef6d6f60abe5d1a74f2deb623777dbe1a9e`
 - Publication receipt SHA-256: `83f5994ef226fb06886e9459086f87973b67167b32e3fbbcb60257b3abf3a83a4`
 - Scored source revision: `1c6420ae9d0f0531edb8c56f7e9e0e586d7c49cd`
 - Publication source revision: `950cd66a4f70cdad5e869294e2179d59d8b6df6a`
 - Report SHA-256: `969fe55a01b218a8bdf111f9bb10f10b64d9409a792caccd23caa802d40a4b79`
 - Scoring receipt SHA-256: `8b17ac5dbcda9befe8ceb720b131497c4d96c7b5f363415309736455b76ee3b3`
 - Provenance SHA-256: `bc97573a2dc17d8b4b5ab2d4ee5144c4fba309be6d42a0e965fb5350c980a94b`
 - Public signer fingerprint: `88e0b36ee7778ab51b82ff9347ed49d8da51ddb22478f093cbdfd0c5f5a7aead`
-

12. Withheld material

The public package intentionally withholds raw prompts and transcripts, scenario and route identifiers, effect-canonicalization and equivalence rules, authority-resolution and denial-propagation internals, private signing keys, provider credentials, and production enforcement implementation details.