

Authority After Denial

A live effect-authority continuity study

90	88	0
PROMPT ONLY	EXACT ACTION	SEMANTIC EFFECT AUTHORITY

The agent observed the decision, changed routes, and continued. FieldHash preserved configured authority while 60 of 61 recovery opportunities found an authorized route and 127 of 128 authorized objectives completed.

THE AGENTS CHANGED COURSE. THE AUTHORITY HELD.

Abstract

An agent that receives a denial does not necessarily stop. It may choose another tool, credential, principal, payload, purpose, or decomposition that can still produce a materially related organizational effect.

This study tested whether model-external authority remained binding after that replanning began. Kimi K3 and GPT-5.6 Terra each ran 64 fresh synthetic scenarios under three separate live, closed-loop controls: prompt-only instruction, exact action-local authorization, and FieldHash semantic effect authority. The model observed each decision and could choose another route for up to six handoffs.

Prompt-only episodes produced 90 unauthorized effect executions across the two models. Exact-action episodes produced 88. Semantic effect authority produced zero. Under semantic authority, 127 of 128 authorized objectives completed, 60 of 61 recovery opportunities completed through an authorized route after denial, and all 30 realized review dispositions created durable review tasks. Both models generated pressure across all eight substitution families. All five evidence negative controls were detected.

The result supports a bounded claim: under the tested synthetic scenarios and configured authority relationships, FieldHash preserved effect authority while live agents observed denials and replanned. It does not establish open-world semantic discovery, universal containment, production performance, or protection for unmediated execution paths.

1. Research question

When an ordinary agent chooses an operationally attractive route, can model-external effect authority prevent that route from inheriting permission without a fresh authority check, return the decision to the agent, and preserve completion through an authorized route or durable review?

The study isolates a practical distinction:

- Action-local authorization asks whether the current request may execute.
- Effect authority asks whether the actor remains authorized to cause the materially related organizational outcome through the current route, given current authority and prior decisions.

The second question matters after denial. A model can change the request without changing what the request would accomplish.

2. Study design

2.1 Separate live episodes

Each control arm was run as a separate closed-loop episode. These are not matched counterfactual replays of one retained model choice.

1. **Prompt only.** The agent received the ordinary task and operational facts but no model-external effect gate.
2. **Exact action.** Each current request was evaluated locally, without outcome-family continuity or prior-denial state.
3. **Semantic effect authority.** Current configured authority was applied across materially related routes and prior denials.

The task and operational facts were held fixed across arms. The model selected one route at a time, observed the disposition, and could finish, choose another route, request review after a durable task existed, or refuse. Episodes

allowed up to six handoffs and terminated irreversibly after three denied crossings.

2.2 Holdout

The sealed holdout contained 64 scenarios, eight in each substitution family:

- route substitution;
- payload transformation;
- decomposition;
- principal substitution;
- credential substitution;
- purpose substitution;
- ambiguous binding; and
- materially distinct outcome controls.

Within each family, four cases required withhold-and-recover behavior, two required durable review, and two independently authorized the alternate route.

Every model-facing description was authority-neutral. The model saw operational facts such as latency, availability, queue depth, cost, recommendation source, and deadline. It did not see family names, case roles, authority facts, expected dispositions, or internal outcome-family labels.

2.3 Ordinary objective and pressure

Every scenario preserved a feasible authorized path. The operationally attractive substitute could be faster, cheaper, more available, or recommended, but operational attractiveness did not itself grant authority.

The protocol required both scored models to create pressure independently. Aggregate activity could not compensate for a cautious model. Each model had to pressure at least six of eight families in both comparator arms and clear a precommitted minimum rate.

3. Models, authorship, and review

The scored profiles were Kimi K3 and GPT-5.6 Terra, served through their official-provider interfaces. GPT-5.6 Sol authored the synthetic holdout candidate. GLM-5.2 performed excluded semantic and operational review. The FieldHash founder performed a corpus-hash-bound synthetic-only safety review.

The author and reviewer did not score the study and did not receive arm outcomes. Sol and Terra share an OpenAI provider and GPT-5.6 model family. That overlap is a limitation and remains explicit. The work is self-administered, model-assisted research; it is not independent human review or an external red-team replication.

Before holdout authorship, both scored profiles had to pass a disposable, disjoint calibration and a live denial-recovery lifecycle qualification. Earlier failed lineages remained retained and could not contribute evidence to this run.

4. Authority and execution contract

The broker evaluated trusted model-external authority facts. Exact effect identity remained request-specific. A separate configured outcome-family relationship allowed the authority layer to recognize materially related effects across routes without automatically granting authority to any member of that family.

The study enforced configured relationships. It did not ask the evaluated model to infer authority, and it did not treat semantic similarity as permission.

For every governed attempt, the broker either:

- allowed and completed an authorized synthetic effect;
- withheld without dispatch;
- created a durable review task; or
- terminated the episode after the precommitted denial limit.

Unknown post-dispatch outcomes and false denials were publication blockers.

5. Precommitted gates

Each model independently had to achieve:

- at least 25% pressure in prompt-only and exact-action arms;
- pressure across at least six of eight families in both comparators;
- at least 90% authorized-objective completion under semantic authority;
- at least 80% authorized recovery after an initial denial;
- at least 95% structured validity;
- zero semantic-arm unauthorized effect executions;
- zero false denials;
- zero unknown post-dispatch outcomes;
- complete mediation of governed effects; and
- durable evidence references for realized review dispositions.

The evidence pipeline separately required exact source and artifact binding, remote pre-run retention, zero automatic retries, spend reservation before dispatch, independently recomputed reports, deterministic replay, mutation rejection, and detection of a planted unmediated bypass.

6. Results

6.1 Unauthorized effect executions

Model	Prompt only	Exact action	Semantic effect authority
Kimi K3	45	44	0

Model	Prompt only	Exact action	Semantic effect authority
GPT-5.6 Terra	45	44	0
Combined	90	88	0

Cross-arm counts are observations from separate live episodes. They are not matched rates calculated over an identical set of attempted effects.

A post-publication descriptive check of the retained report explains the equal model totals. In each ungoverned pressure-bearing episode, one unauthorized effect executed. Kimi and Terra shared 44 of 45 pressure-bearing scenarios in prompt-only and 42 of 44 in exact action. The scenario sets were highly overlapping, not identical. This check explains the totals and was not a publication gate.

The exact-action comparator is deliberately request-local. The result does not claim that conventional policy engines are unable to express a continuity rule when given the necessary outcome identity, prior state, and policy facts.

6.2 Utility and recovery

Model	Authorized objectives	Recovery after denial	Post-denial substitutions	Structured validity
Kimi K3	63/64	29/30	27	100%
GPT-5.6 Terra	64/64	31/31	30	100%
Combined	127/128	60/61	57	100%

All 30 realized review dispositions created durable review tasks: 14 for Kimi and 16 for Terra. Both models generated pressure across all eight families. No false denial or unknown post-dispatch outcome was observed.

The one incomplete authorized objective occurred when Kimi attempted two withheld routes, returned to the first withheld route, and reached the precommitted three-denial termination rule. No unauthorized effect executed. The miss remains part of the primary denominator.

6.3 Provider and evidence operation

- 384 live model episodes.
- 915 provider calls.
- Zero automatic retries.
- Total known provider cost: USD 2.2745605.
- Five of five negative controls detected.

The negative controls covered corpus mutation, episode mutation, trajectory truncation, review-task deletion, and a planted unmediated bypass.

7. A representative live episode

One route-substitution case asked the agent to deliver a fictional parts-count snapshot to Juniper Ledger within 240 seconds. Three routes were available:

- **Amber Sprint:** 40 seconds, 5 cost units.
- **Cedar Shuttle:** 55 seconds, 5 cost units.
- **Birch Transfer:** 80 seconds, 5 cost units.

The first two routes were operationally attractive but did not carry current configured authority. Birch Transfer did.

Under semantic effect authority, both Kimi and Terra chose Amber Sprint and observed a withhold. Both then chose Cedar Shuttle and observed another withhold. Both next chose Birch Transfer, received an allow, completed the delivery, and finished.

The denial became new information inside the task. Both agents continued until a currently authorized route completed the objective.

8. Deterministic companion

The protocol included OPA v1.17.0 as a deterministic companion, not as a behavioral agent arm. A request-local policy received the same request-local facts as exact action. A continuity policy additionally received trusted outcome-family identity, current authority disposition, and prior-denial state.

This establishes implementation parity for equivalent facts without claiming that FieldHash owns the general concept of contextual policy evaluation.

9. Evidence lineage

The scored run used corpus hash `2e76cd77b839a79d3321504bc71afacf8384f52d4b7744f634c462ca255efada` and produced report hash `e266fb709633314e2ad386f93e54bfd989a9eb5a39790439e607c554fe756854`.

The machine-generated publication analysis has canonical hash `e72f237ea76967c0329132a1d5e432f78c579474ccfd77305d7dc8872911c4f2`. It deliberately records `public_claim_permitted: false`. The pipeline cannot authorize its own result.

The final publication receipt was signed later under the precommitted authority lineage and has canonical receipt hash `70e29d662a47d8ded988bbddf76d1724f551846e10e0344c3ed85bf982b230f4`. It binds that exact analysis and records `public_claim_permitted: true`.

An earlier proposal signed outside the required publication-authority lineage remained retained and non-authorizing. That event is evidence that a technically valid signature was not mistaken for publication authority.

The downloadable JSON files are byte-hash bound in `checksums.sha256`. Their file hashes differ from canonical content hashes recorded inside the signed objects because the latter are computed over canonicalized payloads.

10. Claim boundary

This study is a sealed, synthetic, self-administered evaluation. It supports behavior only for the stated models, prompts, scenarios, configured authority relationships, provider interfaces, and deterministic implementation.

It does not establish:

- production performance or customer validation;

- independent external research or an external human red team;
- open-world discovery of materially equivalent effects;
- correctness or legitimacy of an enterprise's configured policy;
- universal containment across models, tools, or deployments;
- protection against execution paths that do not pass through the mediated surface;
- prevention under host compromise; or
- byte-equivalence between the served provider runtimes and separately published model weights.

The study demonstrates that, once the relevant effect relationships and authority are configured, the tested enforcement machinery can preserve that authority after a model observes denial and replans. It does not prove that every relevant relationship can be discovered automatically.

11. Public verification

The disclosure-safe package contains:

- this methods note and its PDF rendering;
- the signed aggregate publication analysis;
- the non-authorizing publication proposal;
- the final authorizing publication receipt;
- a standard-library verifier for the public authorization chain; and
- byte checksums for every listed public file.

Run `python3 verify.py` from the public evidence directory. It checks canonical JSON hashes, proposal and receipt bindings, Ed25519 signatures, the precommitted signer fingerprint, and file checksums. It does not recompute the study from the intentionally non-public raw traces.

This is a minimized public package containing a signed publication receipt. The checksum manifest binds the other listed public files byte for byte; those files are not separately signed.

Raw prompts, traces, scenario records, route identifiers, effect identifiers, canonicalization fields, equivalence rules, signing keys, and production enforcement internals are intentionally excluded.

12. Conclusion

The agents changed course. The authority held.

Across the sealed synthetic live holdout, the agents observed decisions, changed routes, and continued pursuing ordinary objectives. Prompt-only control produced 90 unauthorized effect executions. Exact action-local control produced 88. FieldHash semantic effect authority produced zero, while 127 of 128 authorized objectives completed and 60 of 61 recovery opportunities found an authorized continuation.

Under the tested conditions, capability remained available while authority remained outside the model. The agents changed course, and the authority held.