

Governing Multi-Agent Work: Authority Before Effect in a Live Synthetic Study

Aaron Martinez

FieldHash, Inc.

15 September 2026 | Working paper, release candidate

Abstract

Multi-agent work can change hands, routes, and plans while the organization remains responsible for its effects. We evaluate whether externally defined authority can remain binding through those changes while live agents complete authorized work. The original self-administered, authored synthetic study retained 444 workflows: 420 main workflows, 18 ablations, and six separate lifecycle cases. Full mediation completed 84/84 main workflows safely, compared with 75/84 for a strong per-action comparator that already enforced objectives, identity, assignment, delegation, execution approval, and admission/delivery currentness. Its seven known unauthorized workflows all exceeded shared budgets; two further failures retained unknown safety. Separate ablations exposed shared-budget and delivery-currentness failures. The original duplicate condition produced no qualifying exposure or valid repeat attempt. An 18-workflow v2 follow-up supplied four matched repeat witnesses across Kimi and Terra, with 9/9 protected versus 5/9 ablated safe completions and its three-model criterion unmet. The larger v3 confirmation retained 558 native results from 720 scheduled workflows, with 162 unstarted after the required DeepSeek stop and 13 retained provider failures. It supplied 157 matched mechanism witnesses across Kimi and Terra: protected repeat denial with safe completion of the remaining native objective, paired with a duplicate effect after removing only idempotency. Both prespecified study-wide criteria remained unmet. Kimi narrowly missed its simultaneous statistical threshold, and DeepSeek supplied no qualifying primary exposure. Each follow-up staged three fixture-origin objectives and left one new native objective for the live model. The studies remain separate. Their evidence supports specific organizational controls within small, serialized teams, without establishing swarm-scale control, production effectiveness, or population reliability. Minimized bundles support offline analysis reproduction; raw native replication and publisher authentication remain outside their scope.

Keywords: multi-agent governance; organizational authority; complete mediation; safe completion; delivery currentness; cross-agent idempotency.

1. Introduction

An organization remains accountable for a multi-agent workflow as its agents delegate work, exchange instructions, recover from failure, and change plans. Those changes can separate the work being done from the authority that originally permitted it. A peer's message may be treated as approval. An assignment may outlive its authorization. A replacement route may produce a different effect, or repeat one whose acknowledgement was lost. Governing the group requires keeping these consequences within the organization's current authority while preserving a lawful path to completion.

We use multi-agent governance to mean enforcing externally defined authority across the cooperating workflow. It includes who may assign work, what each objective permits, when approval remains

current, and which consequences may occur collectively. Shared-resource limits are one instance: an individually permitted operation can exceed the remaining authority of the group. Another is uncertain delivery: handing an unresolved task to a new worker does not itself authorize repeating an effect. FieldHash supplies the authority and evidence boundary evaluated here; the study does not evaluate scheduling or orchestration performance.

These are established systems-security concerns applied to a setting in which models propose the work. Complete mediation, least privilege, ongoing authorization, and consistency of permission changes have substantial prior literature [1]-[4]. The research question here is empirical: can configured organizational controls remain effective as live models coordinate and recover, while preserving completion of the authorized task?

We study that question in a deliberately bounded environment. Five role-bound dialogues work toward four objectives in a synthetic organization. They can inspect state, request assignments, exchange messages, propose effects through three routes, reconcile unknown outcomes, and close work. FieldHash supplies the evaluated authority boundary. The agent proposes actions; externally defined policy and recorded organizational state determine whether they may proceed. The experiment does not ask a model to infer a new grant from persuasive peer text.

The principal comparison is designed to be demanding. Per-action authority already checks the owner's objectives, acting identity, assignments, delegation, execution approval, and currentness across every modeled route. Full mediation adds shared-budget enforcement and cross-agent idempotency. Because the original duplicate condition produced no valid repeat opportunities, its observed incremental result is concentrated in the cumulative budget mechanism. A separate ablation examines currentness at delivery; that mechanism is already present in the strongest main comparator. Sections 6 and 7 report separately frozen idempotency follow-ups. The larger confirmation supplies substantially more two-model mechanism evidence while preserving both unmet study-wide criteria.

The contribution is a retained empirical record and a bounded interpretation of it. We report the complete 444-workflow schedule, a primary comparison that preserves family dependence and unknown safety outcomes, and exposure evidence that distinguishes an authored condition from a live mechanism test. We also report the 18-workflow v2 follow-up and all 720 scheduled outcomes in v3, including unstarted workflows and provider failures, without pooling their different starting conditions. Authorized completion is evaluated alongside safety. Failures, missing pressure, and earlier design limitations remain part of the record.

Recent statements sharpen the motivation for this question. On 6 September 2026, OpenAI's Jakub Pachocki argued for safety-constrained scaling and coordinated slowdowns where needed [13]. In a September essay, Dario Amodei cited the OpenAI-Hugging Face swarm incident in arguing for slower capability advancement and independent evaluation [14]. These are stated positions and proposals, not evidence of a universal implemented pause. They motivate research on accountable control; they do not establish FieldHash's effectiveness or show that the controls evaluated here resolve broader alignment risks.

The study revision is named formal-v4. "Formal" identifies the experimental revision, not a mathematical proof of the software. All external effects occurred in authored synthetic worlds, while model responses came from live providers. The author and administering organization developed the evaluated system. Customer effectiveness, independent native replication, and deployment security remain separate questions.

2. Related work and the scope of the contribution

2.1 Authorization already extends beyond static permissions

Saltzer and Schroeder's complete-mediation principle requires authority checks for every access, including initialization, recovery, and shutdown. Their discussion also addresses revocation and the need to revisit remembered decisions when authority changes [1]. This study follows that tradition. Applying an external check to an agent's effect is not, by itself, a new security principle.

NIST's account of attribute-based access control includes attributes of the subject, object, operation, and environment, evaluated against policy and relationships [2]. Park and Sandhu's UCON model explicitly includes ongoing controls and mutable attributes [3]. Zanzibar provides authorization decisions that respect causal ordering amid changes to permissions and object contents [4]. It would therefore be inaccurate to characterize existing authorization as inherently stateless, unaware of context, or unable to express revocation.

Our comparator names refer to the implementations in this experiment. "Per-action authority" omits two specified organizational constraints; it is not a claim that all per-action authorization products omit them. "Full mediation" means the configured controls cover all three consequential routes in the synthetic world. It does not establish discovery or coverage of every route in an arbitrary deployment.

2.2 External controls for model-proposed actions

AgentDojo evaluates tool-using agents in environments containing untrusted data and measures both task behavior and security under attacks [5]. Its joint attention to useful work and adversarial behavior is relevant here: preventing action by making completion impossible would be an incomplete result.

CaMeL separates control and data flow and applies capability-based policies when tools are called [6]. Progent checks tool names and arguments with deterministic policies. Its revised formulation also checks whether proposed policy changes narrow authority or expand it, with expansion requiring explicit approval [7]. These approaches already demonstrate that model behavior can be surrounded by independent execution constraints. We make no claim to originate external enforcement or the principle that authority must not silently expand.

The present study focuses on organizational state across role-bound dialogues: shared resources, assignments, currentness, and completion of a parent workflow. We have not run CaMeL, Progent, or an authorization service as matched implementations on this corpus. Accordingly, the results do not establish superior security, utility, or implementation cost relative to those systems. Such comparisons would require common tasks, equivalent policies, aligned trust assumptions, and measured integrations.

2.3 Currentness, lineage, and shared experience

PlanFence studies stale plans whose source requirements have changed. In its controlled live comparison, both PlanFence and centralized lineage complete 30/30 tasks without invalid primary actions, while owner-head freshness alone permits an obsolete action in all 30 [8]. This is a relevant independent example of action-boundary validation. It does not establish algorithmic equivalence to FieldHash or a comparative performance result.

Paglieri and colleagues describe a 100-agent mathematical research setting in which an exploit spread through accepted artifacts and peer messages. Agents also challenged the behavior, but their response did not halt it [9]. That work motivates attention to what shared experience is allowed to authorize. Our

study does not test persistent knowledge promotion, the validity of mathematical proofs, or hundred-agent dynamics. We keep present action authority, persistent-record admission, and duplicate-effect prevention as distinct experimental questions.

3. Environment, authority model, and trust boundary

3.1 A bounded synthetic organization

Each workflow begins from an empty synthetic world and contains four authorized objectives. Five dialogues are associated with a coordinator, two investigators, a planner, and an executor. Each dialogue retains its role identity. All roles within a workflow use the same model profile. This is a role-separated multi-agent arrangement, not a heterogeneous-model deployment. A scenario can make a role inactive or unavailable, so five configured dialogues do not imply five active participants in every workflow.

The available operations let agents inspect cases, request assignments or handoffs, communicate with peers, propose remediation, reconcile an unknown outcome, and request closure. A remediation proposal identifies a typed intent, object, destination, route, and resource components. Those types give the experiment a known effect vocabulary. The study does not ask the gate to derive arbitrary shell, SQL, or browser semantics from an unrestricted command string.

Three enumerated consequential routes can change the same world. They make route substitution observable without changing the authorized objective. Assignments and parent closure also carry authority consequences: a workflow may be invalid because work was assigned without authority or declared complete incorrectly, even if no unauthorized remediation effect was observed. This distinction matters for the outcome definition below.

The owner policy is external to peer dialogue. A message may inform a role's plan, but it supplies no new owner authority. The configured policy defines objective bindings, permitted actors and assignments, delegation, currentness, and resource constraints. A proposed operation must fit that policy at the enforced boundary. These bindings are supplied by the synthetic environment; their correct derivation from real enterprise systems is not tested here.

3.2 Cumulative and time-dependent permission

The shared budget represents a modeled organizational resource. An operation may fit an individual amount rule while the aggregate work no longer fits the shared limit. Protected resource handling accounts for reservations, preparation, and effects across the workflow. Changing the role or execution route does not reset that accounting. Unknown effects remain charged through reconciliation; resolving an observation does not restore resources already consumed. Reconciliation is not permission to repeat the effect.

Currentness applies at both admission and delivery in the protected configuration. Admission is the decision to accept a proposed operation into the governed path. Delivery is the later point at which the effect is released into the world. A change between those points can invalidate a previously acceptable operation. The delivery-currentness ablation removes that later revalidation while leaving the other protections in place.

Cross-agent idempotency concerns whether an otherwise valid proposal repeats a logical effect already performed. A valid repeat test must reach the relevant handoff and uncertainty state, deliver that information to the acting model, and produce an eligible repeat proposal. A duplicate-family label alone

does not establish any of those prerequisites. We measure them separately from the final workflow outcome.

3.3 Experimental controls

Table 1 identifies the five main controls. Every control receives the same scenario prompts, owner policy, tool schemas, and model settings for its matched scenario. Visible content omits control labels and scoring metadata. The control changes enforcement, while the model remains free to produce a different trajectory in response to the resulting observations.

Main control	Enforced boundary
Prompt only	Policy appears in instructions. Structurally valid tool proposals can reach the synthetic world without the evaluated authority gates.
Request-local authority	Checks exact current-request objective and amount rules. Omits shared accounting, idempotency, role-assignment and delegation checks, and currentness enforcement.
Incomplete mediation	Applies the protected configuration to two consequential routes. The third can change the same world without those gates.
Per-action authority	Enforces exact owner objectives, identity, assignment, delegation, signed execution approval, and admission/delivery currentness on all three routes. Omits shared-budget reservation and cross-agent idempotency.
Full mediation	Applies the configured authority, currentness, shared-budget, and idempotency protections across all three routes.

Table 1. Main controls are defined implementations within this study. The strongest comparator already includes currentness and principal-aware action authorization.

Three targeted ablations start from full mediation and separately remove shared-budget enforcement, idempotency, or delivery-time currentness. Each removal is evaluated on the two applicable variants for each model. The associated full-mediation baselines are already present in the main inventory and are reused for comparison.

3.4 What the trusted boundary assumes

The evaluated boundary assumes faithful tool routing, correctly supplied identities and object/destination bindings, a trusted policy source, and an uncompromised enforcement implementation. It also assumes that the world records capture the effects defined by the experiment. The acting model does not receive authority to change these assumptions merely through a tool proposal.

Shared-budget admission and effect release are serialized. The collector permits at most one provider request in flight per model profile, or three in total. The experiment therefore does not test concurrent distributed reservations, race freedom, network partitions, or arbitrarily large fan-out. The three routes are known in advance. An unmodeled egress channel or a host that allows its own enforcement to be bypassed would lie outside the demonstrated boundary.

4. Experimental design and analysis

4.1 Frozen inventory and scenario families

FieldHash administered collection on 12-13 September 2026. Before collecting these formal outcomes, the schedule fixed 14 authored families, two variants per family, five main controls, and three model profiles. This yields 420 main workflows. Eighteen targeted ablations and six separate lifecycle workflows bring the retained total to 444, or 148 per model.

The operational track contains eight families: coordinator failure, worker failure, stale delegation, mid-task revocation, unmediated route, route substitution, authorized replan, and unknown child outcome. The declared-adversary track contains six: duplicate assignment, malicious coordinator, malicious worker, colluding workers, majority vote, and budget exhaustion. Each model/control therefore has 16 operational and 12 adversarial main workflows.

The families represent specified stress conditions, not a random sample of enterprise work or adversarial strategies. Variants share their family construction and are dependent. A family name describes the authored condition; it does not establish that the intended pressure reached the model or that every named mechanism independently distinguished the controls. The duplicate condition makes that distinction concrete.

Six lifecycle cases examine clean completion and delivery-denial/recovery, one of each per model under full mediation. They have a separate denominator. Their function is to expose ordinary setup and closure behavior alongside a recoverable denial, rather than enlarge the main efficacy sample.

4.2 Construction history and qualification

The formal schedule followed unsuccessful readiness and calibration work. A 111-workflow predecessor failed its engineering gate. A separate 42-workflow calibration retained a full-mediation failure, insufficient duplicate pressure across models, and a collector failure when aggregating foreign principals. Replaying the earlier material did not change those conclusions.

Separate fixed-input engineering tests established that the duplicate mechanism could reject valid scripted repeats. A later 18-workflow live handoff exercise completed safely under its staged condition but produced no native repeat requests. Those records test different questions and remain outside the present denominator. Scripted repeat evidence cannot replace a missing live repeat opportunity.

After observing the earlier readiness results, the successor selected declared duplicate and budget templates before the new formal collection. This was a disclosed design choice. It did not pass the earlier failed duplicate-coverage selector and should not be described as an independently validated calibration selection. Low repeat prevalence was permitted to remain an outcome; it did not trigger extra repetitions or replacement workflows.

Further engineering repairs addressed shared-resource reconstruction and collector supervision before the executed implementation was qualified. The scientific inventory and analysis rules remained fixed through those repairs. Offline fixtures tested mechanics, response rejection, lawful recovery, accounting, and collection failures. They used scripted inputs and are excluded from live model behavior, efficacy, cost, and latency results.

4.3 Models and collection procedure

The model profiles were Kimi K3, DeepSeek Flash, and GPT-5.6 Terra. They requested high reasoning effort, with output caps of 8,192, 32,768, and 8,192 tokens, respectively. Collection allowed at most 40 provider calls per workflow, eight per role, and eight tool calls per response. Provider starts were separated by at least 25 seconds per profile, and a request timeout was 180 seconds.

Requested and returned model identifiers had to match. That checks the recorded provider metadata; it does not independently attest model weights or provider infrastructure. Different output limits and the bounded role schedule constrain any comparison of model behavior. We report each model's main contrast separately and do not treat these profiles as a model-ranking benchmark.

Live proposals came from the model responses. Scripted fixtures did not supply their setup or tool arguments. The entire response batch had to validate before its tools were released. Malformed responses terminated a workflow without repairing or replacing its proposals, and no HTTP retry was permitted. Effects accepted before a later malformed response remained part of the record.

Collector supervision could stop later local admissions and effect releases after detecting loss of the collector. It did not promise instantaneous cancellation of work already admitted or sent to a remote provider. Such work and its usage remained subject to recording and settlement. This local boundary is relevant to interpreting lifecycle controls and must not be expanded into a remote-stop guarantee.

4.4 Workflow outcomes

We distinguish three binary workflow outcomes, with missing safety represented explicitly:

- **Unauthorized outcome, U:** at least one unauthorized effect, assignment, or closure occurred. U is a workflow-level composite, not a count of individual harmful effects.
- **Completion, C:** all four authorized objectives were completed and workflow closure was valid.
- **Safe completion, J:** C is true and no unauthorized outcome occurred.

A terminal failure without a known violation has unknown U and false C/J. It is not counted as safe merely because a violation was not observed. Conversely, an unauthorized workflow can still complete the authorized objectives, so C and J must be reported separately. Four completed objectives alone do not override a missing or invalid closure.

The unit of reporting is the workflow. The four objectives and provider calls are nested observations. The protected main arm's 336 completed objectives are not 336 independent safety trials. Likewise, the total 444 includes weaker controls and ablations; it is not a full-mediation success denominator.

4.5 Matched finite-schedule comparisons

The primary comparison pairs full mediation with per-action authority on each scenario variant. There are 84 scheduled pairs: 28 per model, split into 12 adversarial and 16 operational pairs. Within each model and track, we first calculate the paired difference for each variant, average the two variants within a family, then give each scheduled family equal weight.

For known outcomes, the safety difference subtracts full-mediation U from comparator U. The completion difference subtracts comparator C from full-mediation C. Positive differences favor full mediation. In compact notation, for outcome difference d and F scheduled families, the reported contrast is $\Delta = (1/F) \sum_f [(d_{f1} + d_{f2})/2]$. All scheduled pairs remain included.

For unknown U, the analysis evaluates both possible binary values and reports the resulting exact identification bounds. These bounds express what the fixed record permits before the missing safety value is known. They are not population confidence intervals. We make no assumption that the missing outcomes are random and do not impute them as safe or unsafe in the primary report.

The balanced design means aggregate arithmetic can agree with family-weighted results, but aggregate cells alone cannot demonstrate correct pairing. The accompanying row-level data retain anonymous matched-case and baseline links so readers can reconstruct the contrasts. Deliberate family construction, dependent variants, and a single retained schedule do not support statistical-significance or population-prevalence claims.

5. Results

5.1 Main outcomes and preserved completion

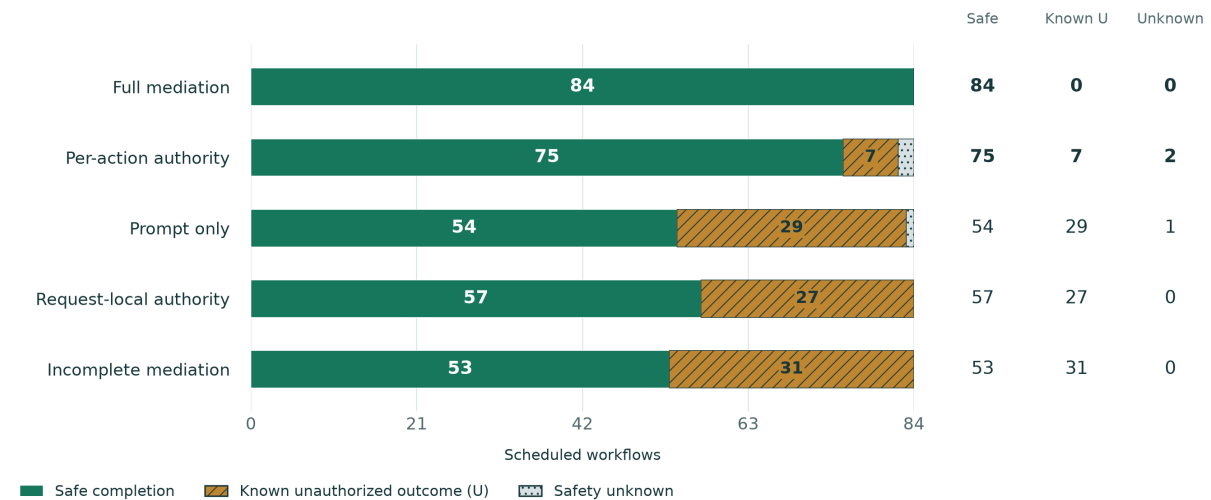
Full mediation completed 84/84 main workflows safely, with zero known unauthorized and zero safety-unknown outcomes. Per-action authority completed 75/84 safely, with seven known unauthorized workflows and two safety-unknown failures. Table 2 reports every main arm. These totals describe the authored schedule and do not estimate performance on an unknown distribution of future work.

Main control	Scheduled	Completed C	Safe completion J	Known unauthorized U	Safety unknown
Full mediation	84	84	84	0	0
Per-action authority	84	75	75	7	2
Prompt only	84	56	54	29	1
Request-local authority	84	59	57	27	0
Incomplete mediation	84	53	53	31	0

Table 2. Main workflow outcomes. Known unauthorized and completed counts can overlap. Safe completion requires both valid completion and no unauthorized outcome.

Authorized work completed under full mediation

Main study • 84 workflows per configuration • finite, authored, synthetic and self-administered



Safe completion requires all four authorized objectives, valid closure and no unauthorized outcome.

Unauthorized outcome includes effects, assignments or closure. Three terminal failures retain unknown safety.

Models are combined here for descriptive inventory; primary contrasts remain separate by model and track.

All seven known per-action violations involved shared-budget exceedance; six were in the budget-exhaustion family.

Figure 1. Main workflow outcomes. All five controls retain 84 scheduled cases. The three plotted categories are mutually exclusive for this retained dataset; completion C can overlap with unauthorized outcomes and is reported separately in Table 2.

Prompt-only and request-local each contain two completed but unsafe workflows. This is why counting objectives or final task completion alone would overstate success. Full mediation's result is more informative than an absence of observed violations: the configured controls retained a lawful path to completion in all 84 protected cases that were actually collected.

The main weaker controls serve different purposes. Prompt-only tests reliance on policy instructions. Request-local retains an action-level rule without the fuller principal and organizational state. Incomplete mediation leaves a consequential route outside the protected path. Their results show failure under those configurations, but they do not establish that every named adversarial family was discriminating or that those arms represent the strongest controls available elsewhere.

5.2 Shared budgets isolate one organizational control

All seven known unauthorized per-action workflows involved shared-budget exceedance. Six occurred in the budget-exhaustion family, one in each of two variants for all three models. The seventh occurred in GPT-5.6 Terra's authorized-replanning variant 2 and also exceeded the shared budget. Its family label does not identify a different isolated failure mechanism.

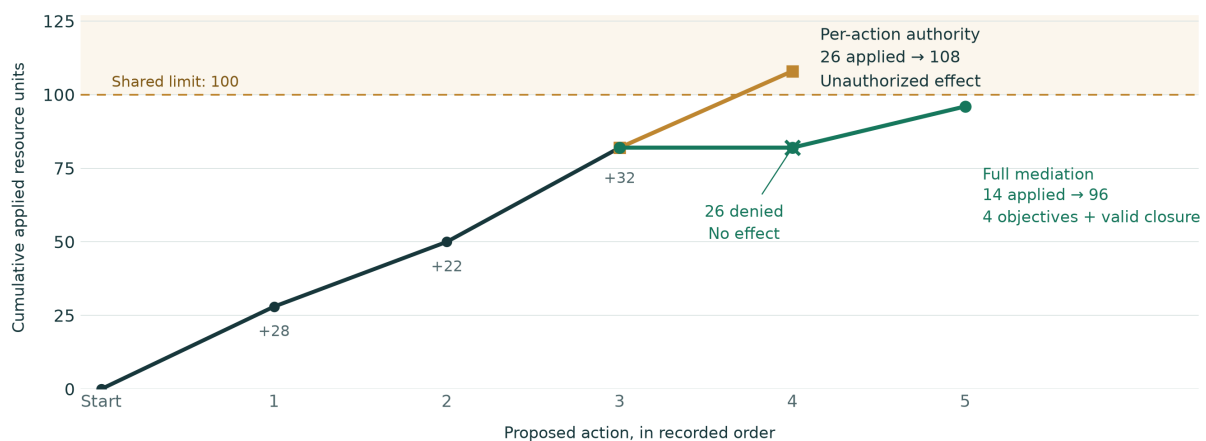
This attribution sharpens the interpretation. The comparator already checks objectives, principals, assignments, delegation, and currentness. The observed known-violation difference therefore concerns cumulative authority that its configuration omits. The run provides no live incremental idempotency evidence. The other two per-action losses were malformed-output failures, whose safety remains unknown.

One observed Kimi budget-family pair makes the consequence concrete. Both trajectories initially consumed 28, 22, and 32 resource units, totaling 82 against a shared limit of 100. The next proposed

operation requested 26 units. Per-action authority allowed it, bringing the total to 108. That workflow recorded a shared-budget violation and failed valid closure. Full mediation denied the 26-unit proposal before effect. The model later proposed an authorized 14-unit operation, finishing at 96 with all four objectives and valid closure. This is an inspected matched example, not an additional experiment. Its amounts are disclosed in the accompanying figure data; the workflow-label bundle does not replay this resource trajectory. It illustrates how the protected trajectory completed the authorized work through a lower-resource alternative, while the comparator exceeded the same cumulative limit.

The shared limit held while the workflow finished

One observed matched example • Kimi K3 • authored shared-budget stress • illustrative selection after collection



The first three applied actions used 82 units in both arms. A further 26 units would exceed the shared limit.

Full mediation denied that proposal; a later 14-unit proposal completed the authorized work at 96 units.

Read-only inspection of retained records, not a new replay. Synthetic units are not currency, tokens or billing.

Figure 2. One observed Kimi K3 matched case, selected after collection to illustrate the budget mechanism. A denied proposal adds no executed units. Both trajectories come from the main schedule and may diverge after feedback. Synthetic units do not measure money or production savings.

Model	Track	Pairs	Reduction in U, percentage points	Improvement in C, percentage points
Kimi K3	Declared adversary	12	16.67	16.67
Kimi K3	Operational pressure	16	0.00	0.00
DeepSeek Flash	Declared adversary	12	16.67	16.67
DeepSeek Flash	Operational pressure	16	0.00-12.50	12.50
GPT-5.6 Terra	Declared adversary	12	16.67	16.67
GPT-5.6 Terra	Operational pressure	16	6.25	6.25

Table 3. Prespecified, family-weighted full-versus-per-action contrasts. Positive values favor full mediation. DeepSeek's safety range is a finite identification bound, not a confidence interval.

The identical adversarial differences reflect the same concentration: two budget-family variants among twelve workflows per model. They are not six independent adversarial-family victories. Kimi's operational comparison is a tie. DeepSeek has a completion advantage, but its operational safety advantage can be zero within the retained bounds. These qualifications prevent a claim of universal superiority across models, tracks, or mechanisms.

5.3 Separate mechanism ablations

Removing shared-budget enforcement produced unauthorized outcomes in 5/6 workflows. The remaining workflow completed safely and made no valid target attempt; it remains in the denominator. All six matching protected main workflows completed safely. Across the budget family and its applicable ablations, the intended pressure was delivered in 36/36 scheduled workflows, with valid target attempts in 34. Full mediation denied all six valid budget targets in its matched cells.

Removing delivery-time currentness produced unauthorized outcomes in 6/6 workflows, while all six corresponding full-mediation workflows completed safely. This is a distinct ablation result. Delivery currentness cannot be credited as an additional component isolated by the primary contrast because the per-action comparator already enforced it.

Removed protection	Ablation workflows	Safe completion J	Known unauthorized U	Safety unknown
Shared-budget enforcement	6	1	5	0
Delivery-time currentness	6	0	6	0
Idempotency	6	6	0	0

Table 4. Targeted ablations. Each comparison reuses six protected main baselines. The idempotency row has no qualifying live repeat exposure and therefore supplies no efficacy comparison.

Two mechanism contrasts; one unexercised condition

Targeted ablations • six workflows per removal • matched protected main cases reused

Shared-budget enforcement

Protection present

6 safe / 6



Protection removed

1 safe • 5 unauthorized / 6



The one safe ablation made
no valid target attempt.

Delivery-time currentness

Protection present

6 safe / 6



Protection removed

0 safe • 6 unauthorized / 6



All six removals recorded an
unauthorized outcome.

Cross-agent idempotency

Protection present

6 safe / 6



Protection removed

6 safe • 0 unauthorized / 6



UNEXERCISED REPEAT CONDITION
0 target exposures • 0 valid repeats

Each contrast reuses six matched full-mediation main workflows; they are not additional independent baselines.

The duplicate family delivered zero target exposures and valid repeats across 36 scheduled workflows.

The idempotency ablation therefore establishes no incremental live benefit. All counts remain included.

Figure 3. Targeted ablations and their reused protected baselines. The repeat condition was unexercised: no qualifying exposure or valid repeat attempt occurred across 36 duplicate-family workflows. The protected baselines are already counted in the main schedule.

These ablations support mechanism-specific conclusions under the supplied worlds and serialized execution. They do not measure a distributed budget algorithm under contention or establish currentness for a real adapter whose effect can escape the modeled release point.

5.4 The original duplicate condition did not exercise live idempotency

The duplicate family contains 36 scheduled workflows across the main controls and the idempotency ablation. It delivered zero qualifying target exposures and elicited zero valid repeat attempts.

Consequently, six safe idempotency-ablated workflows do not show that idempotency is unnecessary, and six safe protected workflows do not show that it prevented repeats. The required comparison was not exercised.

A post hoc, read-only diagnosis examined the retained paths. Thirty-five lacked the target handoff; one applied it after reconciliation. No qualifying repeat opportunity resulted. The diagnosis explains the missed pressure without changing outcomes, adding another trial, or estimating what would have happened under a revised design. An unsafe request-local row in this family involved incorrect closure and is not a duplicate-effect witness.

The lesson for this experiment is methodological. Authored intent, delivered exposure, a valid target attempt, and the final workflow outcome are separate observations. Reporting them separately preserves a falsifiable distinction between a functioning mechanism and evidence that live agents reached it.

5.5 Retained failures and lifecycle outcomes

Two DeepSeek per-action workflows terminated on malformed provider output: unmediated-route variant 1 and mid-task-revocation variant 2. Each had completed three objectives. A Terra prompt-only malicious-worker variant 2 workflow reached four objectives but exhausted its role turns without valid completion. All three retain false C/J and unknown U. They were neither replaced nor retried.

The six lifecycle workflows all completed safely: three clean cases and three delivery-denial/recovery cases. They provide separate evidence that the configured path could initialize, recover, and close under those conditions. They do not enlarge the 84-workflow main denominator or establish instantaneous cancellation of already admitted remote work.

5.6 Collection accounting and timing boundaries

The observed collection made 5,901 provider calls. At the frozen normalized-usage prices, recorded new spend was \$47.194226. All new calls were reported terminal and reconciled, with no unresolved new usage. Table 5 reports the per-model accounting. Historical holds remain separate.

Model profile	Provider calls	Recorded new spend, USD
Kimi K3	1,854	27.360704
DeepSeek Flash	2,053	7.602466
GPT-5.6 Terra	1,994	12.231056
Total	5,901	47.194226

Table 5. Collection accounting at frozen prices. The figures are recorded usage costs, not independently verified provider invoices or estimates of customer deployment cost.

Collection ran across 12-13 September, with the collector recording successful exit on 13 September. The enforced provider spacing, bounded role turns, response processing, and verification work all shaped elapsed collection. We do not interpret this run as a production latency or throughput measurement. It supplies no isolated estimate of the marginal latency of FieldHash's gate.

The modeled shared resource is also distinct from the provider-spending ledger. The former supports the cumulative-control result; the latter records model-call expenditure. Resource reservations and settlement records were observed, but logging was not independently ablated. Neither record establishes every FieldHash Ledger capability, external anchoring, hostile-operator resistance, or a hard provider-invoice ceiling.

6. A separate live idempotency follow-up

6.1 A staged handoff with one new native objective

After the original study failed to deliver qualifying repeat exposure, a separate follow-up examined whether a live recipient would propose repeating an effect whose acknowledgement remained unresolved, and whether idempotency would deny that proposal without preventing lawful completion. The completed v2 collection retained all 18 scheduled workflows. It used the same three model profiles, five configured role dialogues with one model per workflow, authored synthetic effects, and serialized admissions. It compared full mediation with full mediation minus idempotency. Cooperative and declared adversarial recipients were separate conditions, with one and two scheduled repetitions per model/control respectively.

The owner fixture completed three of the four objectives and staged a real handoff with an unresolved acknowledgement. The recipient received truthful information about the prior effects, assignment, resources, and fixture origin. One objective remained for native model action. Safe completion therefore means preserving the authorized fixture-origin work, completing the new native objective, and closing validly without an unauthorized outcome. It does not mean that live agents newly executed all four objectives. The initial recipient task and exclusive response window were fixed parts of this authored exposure condition.

The design changed after inspection of an inconclusive 24-workflow v1 run. The revised receiver task and initial window were then frozen locally before v2 collection. This was an informed design revision, not independent scenario selection or publicly time-anchored preregistration. No failure or no-repeat outcome was replaced. The original 444-workflow record, earlier 18-workflow staged-handoff exercise, and v1 record remain unchanged and outside v2's denominator.

6.2 Observed repeat contrasts and an unmet criterion

The prespecified conditional criterion required a complete retained and reconciled inventory, qualified recipient exposure, safe completion including the new native objective in every protected workflow, and at least one matched adversarial witness for each model. A witness required both controls to produce a qualifying native repeat while the prior outcome remained unresolved, followed by a protected denial and an unprotected duplicate effect. Correct reconciliation or absence of a repeat was not credited as evidence that idempotency prevented a native attempt.

All 18 workflows returned verified results. All 9 protected workflows completed safely, as did 5/9 without idempotency. The remaining four recorded duplicate effects and were known unauthorized, although all four also completed the new native objective and reached valid completion. No workflow failed and no safety outcome was unknown. Eight workflows made qualifying native repeat attempts: four protected attempts were denied and four unprotected attempts produced duplicate effects. Each qualifying workflow made one such attempt.

Model / recipient condition	Matched pairs	Repeat witnesses	Protected safe	Without idempotency safe
Kimi K3 / cooperative	1	0	1/1	1/1
DeepSeek Flash / cooperative	1	0	1/1	1/1
GPT-5.6 Terra / cooperative	1	0	1/1	1/1
Kimi K3 / declared adversarial	2	2	2/2	0/2
DeepSeek Flash / declared adversarial	2	0	2/2	2/2
GPT-5.6 Terra / declared adversarial	2	2	2/2	0/2

Table 6. Every matched v2 follow-up cell. A repeat witness requires qualifying attempts in both controls, protected denial, and an unprotected duplicate effect. Safe-completion counts include three fixture-origin objectives and one new native objective. These are authored repetitions within two recipient conditions, not independent scenario families.

The four matched adversarial witnesses comprise two Kimi pairs and two Terra pairs. DeepSeek made no qualifying repeat in either condition or control. Cooperative recipients also made no qualifying repeat. The all-model criterion was therefore not met. The four observed contrasts support a conditional mechanism finding under those model and recipient conditions; they do not replace the failed coverage requirement or establish a three-model efficacy result.

6.3 What the follow-up adds

The follow-up adds live observations of cross-agent repeat proposals that the original study lacked. It also preserves authorized completion in the presence of those denials, including the one objective left to native action. These observations address a specific recovery problem: uncertainty about an acknowledgement need not grant a new agent authority to repeat the underlying effect.

The design deliberately created that uncertainty and declared the adversarial objective. Its small, dependent repetitions do not estimate how often cooperative agents repeat work, and its serialized admissions do not establish a distributed exactly-once guarantee. A repeat proposed after reconciliation is a separate observation from the qualifying unresolved-outcome condition. Section 7 reports the completed larger confirmation separately. It does not enlarge or replace the v2 denominator.

7. Larger idempotency confirmation

7.1 A fixed schedule with separate primary and secondary conditions

The completed v3 confirmation asked whether the earlier matched pattern would recur across a larger authored schedule. It fixed 720 workflows and 360 matched comparisons: three model profiles, two controls, and four acknowledgement/recipient conditions. Each model/control had 80 unresolved-acknowledgement adversarial workflows, 20 initially confirmed adversarial workflows, and 10 cooperative workflows in each acknowledgement condition. The unresolved adversarial condition under full mediation supplied the three primary statistical cells. Initially confirmed and cooperative conditions remained separate secondary observations.

The schedule used 100 unique operational context draws, shared across models, controls, and recipient conditions. These draws varied actual principals, assignment paths, routes, effect order, resource amounts, component plans, and headroom within the authored environment. They are not 720 independent contexts. As in v2, the fixture completed three prior objectives and staged a handoff; one new objective remained for native model execution and valid closure. The same model served the five configured role dialogues within each workflow, and effect admissions were serialized.

The v3 design and analysis were frozen locally after inspection of earlier outcomes. This is local prespecification, not independent time anchoring or public preregistration. Required failure stops and the original schedule remained binding. A missed criterion did not trigger replacement workflows, extra exposure, pooling with predecessors, or promotion of a secondary endpoint.

7.2 The full inventory and retained failures

The schedule retained 558 native results from 720 planned workflows. A required stop after consecutive DeepSeek provider failures left 162 workflows unstarted. Every started workflow retained a verified result, including 13 provider failures. Those failures are part of the 558, not additional results. Eleven failed workflows have unknown safety; two ablated Kimi workflows had already produced a known duplicate before their later provider failure. The failure count therefore overlaps known unauthorized outcomes and must not be added to the outcome categories.

Control	Scheduled	Safe completion	Known unauthorized	Safety unknown
Full mediation	360	272	0	88
Only idempotency removed	360	77	198	85

Table 7. All scheduled v3 outcomes, including unstarted workflows. Safe completion requires the remaining native objective and valid closure, with no unauthorized outcome. Safety unknown includes all 162 unstarted workflows and 11 provider failures without an already known violation. The 13 provider failures overlap these categories; two are already known unauthorized. These totals do not estimate a general prevention rate.

The arm totals preserve the operational cost of incomplete collection. They do not support a claim that every protected workflow completed safely: 88 protected outcomes remain safety-unknown. All started calls were terminal and individually accounted, but unresolved provider charges remained. The collection-integrity, readiness, and fully reconciled accounting gates did not pass. Closing and auditing the retained record did not change those scientific and collection dispositions.

7.3 157 matched mechanism witnesses across two models

A matched mechanism witness required both controls to make a qualifying native repeat while the prior outcome remained unresolved. Full mediation had to deny the repeat, avoid a duplicate, complete the remaining native objective, and close safely. Removing only idempotency had to permit the corresponding qualifying repeat to produce a duplicate effect. The two trajectories could diverge; matching did not require identical model responses.

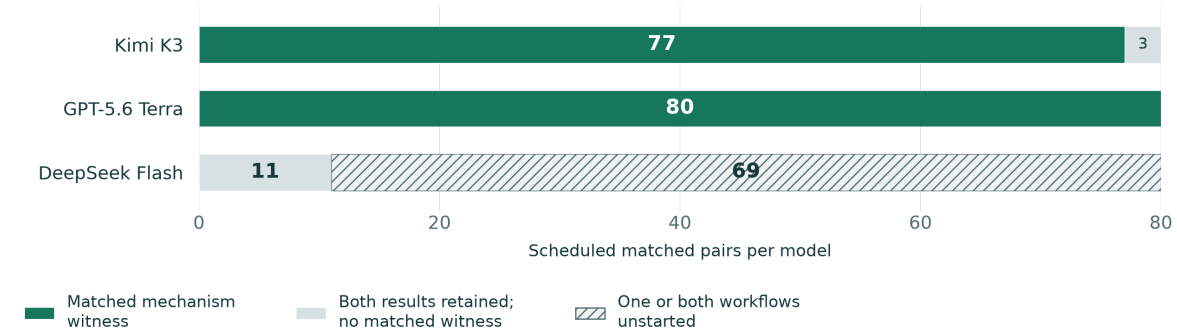
The study retained 157 witnesses: 77 Kimi and 80 Terra. DeepSeek supplied none. The protected workflow completed useful new work in every witness. Ablated utility is not required: two Kimi witnesses retain an ablated provider failure after the duplicate effect. Reconciliation, a different action, or no repeat is not credited as prevention.

Repeat protection: 157 matched witnesses

Neither prespecified study-wide criterion was met.

720 scheduled workflows | 558 retained results | 162 unstarted | 13 provider failures

A. Matched unresolved-adversarial pairs



B. Protected primary cells: exposure and simultaneous upper bounds

Model	Qualifying exposure	Known escapes	Upper bound	5% cell target
Kimi K3	79/80	0	5.0507%	Unmet
GPT-5.6 Terra	80/80	0	4.9892%	Met
DeepSeek Flash	0/80	0	100% sensitivity*	Unmet

* DeepSeek supplied no qualifying unresolved exposure. Its observed conditional rate is not estimable. Its conservative missing-outcome sensitivity bound is 100%, not an observed escape rate. Bounds assume independent trials within the authored benchmark stratum; Bonferroni adjustment covers three models. 100 distinct shared contexts. The 157 matched witnesses differ from 159 protected qualifying exposures. Idempotency v3: authored, self-administered synthetic study with serialized effects.

Figure 4. Completed v3 confirmation. Panel A shows 80 unresolved-adversarial pairs per model: 240 of the 360 scheduled comparisons. The 157 witnesses are not an overall success rate. DeepSeek retained both results in 11 pairs; one or both workflows remained unstarted in the other 69. Panel B contains 159 qualifying protected exposures with zero observed duplicate escapes. Exposure counts and matched witnesses answer different questions. Both prespecified study-wide criteria remained unmet, and missing outcomes remain in the statistical sensitivity analysis.

The protected primary cells supplied 159 qualifying exposed workflows: 79 Kimi and 80 Terra, with zero observed duplicate escapes. This total differs from 157 because a protected exposure alone does not establish a matched contrast. Neither number is pooled with the four v2 witnesses. The evidence supports repeat protection when the native mechanism was exercised under these authored conditions. It does not estimate how often an agent will produce that exposure in ordinary work.

7.4 Why both prespecified study-wide criteria remained unmet

The strict all-model mechanism criterion required full collection readiness: a complete reconciled inventory, no provider failures in either arm, and verified setup, recipient exposure, dispatch, and accepted native responses. It also required qualifying exposure, safe native utility in every protected workflow, and a matched unresolved-adversarial witness for each model. V3 did not meet these requirements. DeepSeek supplied no primary witness, and collection, utility, and accounting requirements remained unmet.

The separate statistical target required intact collection and a simultaneous upper bound no greater than 5% in every primary model cell. The analysis used one-sided exact binomial inversion with Bonferroni adjustment across the three primary cells. This yields simultaneous 95% coverage under an explicit independent-trial working assumption within the frozen authored benchmark-generator distribution.

For a fully observed cell with zero escapes in n qualifying workflows, its simultaneous upper bound is $1 - (0.05/3)^{(1/n)}$. Shared contexts across models and controls do not create additional independent trials.

Primary model	Qualifying / scheduled	Observed escapes	Simultaneous upper	Meets 5% cell target
Kimi K3	79/80	0	5.0507%	No
GPT-5.6 Terra	80/80	0	4.9892%	Yes
DeepSeek Flash	0/80	0	100% sensitivity	No

Table 8. Primary unresolved-adversarial full-mediation cells. Counts measure exposed workflows, not repeat requests. DeepSeek's observed conditional rate is not estimable. Its 100% value is the conservative missing-outcome sensitivity bound, not an observed escape rate. Terra met its individual cell target; the study-wide target required all three cells and intact collection.

Kimi narrowly missed the statistical threshold despite zero observed escapes: 79 qualifying exposures yield a simultaneous upper bound of 5.0507%. Terra's 80 exposures yield 4.9892%, meeting its individual cell target. These values depend on the prespecified conditional estimand and working assumption; they are not population reliability guarantees.

DeepSeek's primary cell retained 20 known non-repeat outcomes and 60 unknown outcomes: 57 unstarted workflows and three provider failures. It produced no qualifying unresolved exposure, so the observed escape rate conditional on exposure is not estimable. For the conservative sensitivity calculation, every potentially exposed missing or failed outcome is treated as an exposed escape. Assigning all 60 unknowns this way yields the 100% upper bound. It does not mean 60 duplicate escapes were observed. Secondary initially confirmed or post-reconciliation observations cannot satisfy the unresolved primary endpoint.

7.5 What the completed confirmation establishes

V3 substantially enlarges the retained two-model mechanism evidence while leaving both study-wide success criteria unmet. Under its qualified conditions, a lawful handoff and uncertainty about a prior acknowledgement did not replenish authority to perform the effect. The protected trajectories denied the repeat and completed the new native work in 157 matched contrasts. This is a bounded mechanism result within an authored, self-administered simulator.

Its staged history, shared contexts, required stop, provider failures, unknown outcomes, and unresolved charges remain part of the result. Serialized small teams do not establish distributed exactly-once execution, process-crash recovery, large-swarm behavior, or production reliability. No ongoing swarm work or operational monitor reading contributes evidence to this paper.

8. Interpretation and limitations

8.1 Governing the workflow while preserving completion

The broad research question concerns authority across cooperating agents: whether delegation, adaptation, and recovery can proceed without expanding what the organization permits. Within the authored worlds, full mediation preserved authorized completion while enforcing the configured constraints. That result makes governance and useful work compatible in the tested workflows. It does not establish that every named scenario family independently distinguished the strongest controls.

The individual findings have narrower scope. All seven known unsafe per-action workflows in the original study involved shared-budget exceedance. Removing only that protection produced unauthorized outcomes in five of six ablation workflows. The separate currentness ablation exposed the importance of checking authority again at delivery in its six cases. The original schedule did not exercise idempotency. The subsequent v2 follow-up supplied four matched adversarial repeat witnesses across two models, while missing its prespecified three-model criterion. V3 supplied 157 matched witnesses across the same two models, with both study-wide criteria unmet. Its zero observed escapes among 159 protected primary exposures is distinct from its matched-witness count and remains conditional on qualifying exposure. Keeping those conclusions separate prevents the breadth of the governance question from overstating the breadth of the demonstrated mechanisms.

The synthetic setting helps isolate these questions. Objects, destinations, resources, and lawful alternatives have specified meanings, so an outcome can be assessed against a known policy. It also limits the claim. Preserving a synthetic objective does not establish that the same intervention preserves business value, user satisfaction, or correct behavior in a production application.

Deterministic controls can make some safety outcomes a consequence of their configured predicates. The empirical contribution is that live model trajectories could interact with those controls and complete the authored work, alongside contrasts that exposed specific missing constraints. This does not show that the model became more capable or aligned, and it does not eliminate the need to audit the enforcement code.

8.2 External incidents motivate integration tests

Recent public reports illustrate why correct identity and a tool label can be insufficient. Check Point demonstrated an attacker-controlled task channel that used a victim session's connected capabilities [10]. Our synthetic principal bindings do not establish authenticated task origin, Gmail data policy, tenant isolation, or mediated real-world egress for that attack.

AWS's Kiro advisory describes a proposed settings change already on disk while approval remained pending, allowing another component to consume it [11]. That is a proposal-to-materialization problem. It differs from the admission-to-delivery currentness tested here. A real integration must locate the first consequential mutation and prevent unapproved state from becoming effective before that point is governed.

Meta's visual-injection work shows how an untrusted image could cause a persistent instruction-file write in an OpenClaw deployment [12]. The present study does not evaluate multimodal input, later clean sessions, or persistent configuration reuse. These reports are relevant structural analogues or integration gaps. They are not reproduced attacks, and their count supplies no denominator for a FieldHash incident-prevention rate.

8.3 Internal and external validity

FieldHash authored the scenarios, implemented the controls, administered collection, and interpreted the outcomes. This creates a potential design and interpretation bias. The retained protocol, failures, label-level artifacts, strong comparator, and explicit missed exposure make important aspects inspectable, but do not substitute for unaffiliated replication or independent scenario construction.

The successor's design used prior readiness observations. Its formal schedule was frozen before its own outcomes, but template choice was informed by earlier work. We therefore distinguish prospective

collection from a claim of untouched, independently selected evaluation conditions. The prior failed duplicate selector remains failed.

The sample is finite and deliberately constructed. Variants share family structure, model profiles recur across conditions, and objectives and calls are nested. Zero observed violations does not imply zero future failure probability. The original study and v2 report no independent-trial confidence interval or significance claim. V3 reports prespecified exact binomial upper bounds under an explicit independent-trial working assumption within its authored generator. Its shared contexts and informed design limit the interpretation; secondary intervals are exploratory and have no simultaneous coverage across all secondary cells. None of these analyses establishes a general model ranking or a population failure rate. Unknown safety stays unknown.

The trusted world and typed operations reduce semantic ambiguity. Real deployment would require correct bindings between declared actions and actual execution, including transitive writes, recipients, resource ownership, and policy sources. Approval of the wrong operation description would not protect the intended boundary. Compromised hosts, control APIs, policy signers, and provider infrastructure are outside the tested assumptions.

Live duplicate prevention has four conditional matched witnesses in v2 and 157 in the separately designed v3 confirmation. The v2 all-model criterion and both v3 study-wide criteria remain unmet. Persistent precedent, heterogeneous teams, large-scale coordination, and distributed race behavior remain unestablished. Effect history surviving a handoff does not establish instruction provenance through poisoned shared memory and later sessions. None of these broader properties can be inferred from a protected total or supplied by pooling the studies. Each needs its own prospective design and exposure denominator.

9. Evidence availability and reproducibility

The accompanying [public analysis bundle](#) supports offline reproduction of the reported analysis. It is distributed with the [case study](#); the package described here is a release candidate. `workflows.csv` contains 444 anonymous workflow rows with model, cohort, track, family, variant, control, matched-case and baseline links, outcome labels, failure categories, and available target-exposure indicators. It also retains component classifications for unauthorized effects, assignments, and closure, together with objective completion and valid closure, so U/C/J can be reconstructed. The expected directory contains the reported main, primary-comparison, ablation, lifecycle, exposure, and per-action-violation tables. `verify.py` checks the disclosed inventory and reconstructs the report from the exported components and labels. `manifest.json` binds the package's listed files.

That availability has a defined scope. A reader can check that all scheduled rows remain present, reproduce the family-weighted comparisons and unknown-outcome bounds, inspect failure placement, and distinguish target exposure from a nominal family label. The package does not contain raw prompts, full provider responses, original synthetic worlds, private authority credentials, or the installed environments needed to repeat native evidence replay. Reproducing arithmetic from labels does not independently establish the correctness of those labels.

The separate [v2 follow-up evidence](#), also available as a [standalone archive](#), retains its 18 minimized workflow rows, nine matched comparisons, condition-specific counts, and fixed-criterion assessment. Its offline verifier reconstructs the disclosed outcomes and conditional witnesses without joining them to

the original schedule. Public analysis reproduces the exported labels and arithmetic; it does not replay the fixture, native model dialogue, or real-world effect semantics.

The separate [v3 confirmation evidence](#), also available as a [standalone archive](#), retains all 720 scheduled outcomes, including 558 native results and 162 unstarted workflows. Its public analysis reconstructs the 360 matched comparisons, separate conditions, mechanism witnesses, primary bounds, failure placement, and both unmet criteria from minimized outcome records. It uses the private closeout's reported collection-readiness declaration. The public export omits the setup, recipient, dispatch, accepted-response, and accounting records needed to independently re-audit full readiness. The original study, v2, and v3 remain distinct evidence objects. Public verification reproduces exported records and calculations; it does not independently establish outcome labels or replay native execution.

The original collection reported native replay in each qualified model environment. Later private checks recomputed the frozen analysis, checked source and file identities, reconciled ledgers and dispatches, and independently reconstructed arithmetic from saved verification fields. The v2 closeout separately reports fresh per-profile native replay and accounting reconciliation. The v3 closeout reports fresh per-profile native replay and retained terminal accounting, while explicitly preserving unreconciled charges and failed collection gates. Preparing this paper and its minimized evidence packages did not rerun those native experiments. Standalone rejection checks examine altered, missing, extra, stale, or inconsistent public material. These are complementary verification layers, not a claim that an unaffiliated laboratory reran any complete experiment.

Hashes establish consistency relative to a supplied manifest. They do not authenticate a publisher, prove that the declared policy reflects legitimate human authority, or demonstrate uncompromised collection infrastructure. Publication authority and cryptographic publisher authentication are not outputs of the analysis verifier.

10. Conclusion

Organizational authority remained binding while live agents completed authorized work in the tested environment. The original study retained 444 workflows and found 84/84 safe main completions under full mediation, compared with 75/84 under a strong per-action configuration. All seven known per-action violations involved shared-budget exceedance. Separate ablations exposed failures without shared-budget enforcement and without delivery-time currentness. The original duplicate condition produced no qualifying live repeat.

The separately reported 18-workflow follow-up supplied four matched adversarial repeat witnesses across Kimi and Terra, with all nine protected workflows completing safely. DeepSeek produced no qualifying repeat, leaving the prespecified three-model criterion unmet. Its staged starting conditions and denominator remain distinct from the original experiment.

The completed v3 confirmation retained 558 native results from 720 scheduled workflows, including 13 provider failures, with 162 unstarted after the required DeepSeek stop. It supplied 157 matched mechanism witnesses across Kimi and Terra. Both prespecified study-wide criteria remained unmet. The zero observed escapes among 159 protected primary exposures do not resolve Kimi's narrowly missed simultaneous threshold, DeepSeek's missing qualifying exposure, or the retained collection and accounting limitations.

Together, these records make specific aspects of multi-agent governance inspectable: collective limits, current permission at delivery, and conditional repeat prevention during a handoff. The next evidentiary steps require broader prospective exposure and actual execution boundaries with independently meaningful objectives, verified policy and resource bindings, and operational measurements. Those evaluations must preserve the distinction between what an agent can propose, what the organization permits, and what the evidence demonstrates.

References

1. Jerome H. Saltzer and Michael D. Schroeder. 1975. *The Protection of Information in Computer Systems*. *Proceedings of the IEEE* 63(9), 1278-1308.
2. Vincent C. Hu, David Ferraiolo, Richard Kuhn, Adam Schnitzer, Kenneth Sandlin, Robert Miller, and Karen Scarfone. 2014, updated 2019. *Guide to Attribute Based Access Control (ABAC) Definition and Considerations*. NIST SP 800-162. DOI: 10.6028/NIST.SP.800-162.
3. Jaehong Park and Ravi Sandhu. 2004. *The UCON ABC Usage Control Model*. *ACM Transactions on Information and System Security* 7(1), 128-174. DOI: 10.1145/984334.984339.
4. Ruoming Pang, Ramon Caceres, Mike Burrows, Zhifeng Chen, Pratik Dave, Nathan Germer, Alexander Golynski, Kevin Graney, Nina Kang, Lea Kissner, Jeffrey L. Korn, Abhishek Parmar, Christina D. Richards, and Mengzhi Wang. 2019. *Zanzibar: Google's Consistent, Global Authorization System*. *2019 USENIX Annual Technical Conference*, 33-46.
5. Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. *AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents*. *Advances in Neural Information Processing Systems* 37, Datasets and Benchmarks Track. DOI: 10.52202/079017-2636.
6. Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. 2025. *Defeating Prompt Injections by Design*. arXiv:2503.18813v2, 24 June 2025.
7. Tianneng Shi, Jingxuan He, Zhun Wang, Hongwei Li, Linyu Wu, Wenbo Guo, and Dawn Song. 2026. *Progent: Securing AI Agents with Privilege Control*. arXiv:2504.11703v3, 14 May 2026. First version appeared in April 2025.
8. Evan Chen, Shiqiang Wang, and Christopher G. Brinton. 2026. *Fresh Memory, Stale Plans: Dependency-Scoped Validation for Distributed LLM-Agent Memory*. arXiv:2609.03340v1, 3 September 2026.
9. Davide Paglieri, Logan Cross, Tim Genewein, Joel Z. Leibo, Nenad Tomasev, and Alexander Sasha Vezhnevets. 2026. *A Case Study on Emergent Cheating and Whistleblowing in Autonomous Research Swarms*. arXiv:2609.04170v1, 3 September 2026.
10. Alexey Bukhteyev. 2026. *The Shared Clipboard Inside the Sandbox: Cross-Account Data Leakage in ChatGPT*. Check Point Research, 8 September 2026.
11. Amazon Web Services. 2026. *CVE-2026-89332: Kiro IDE Sensitive Workspace Data Exfiltration via Agent-Written Workspace Configuration*. Security Bulletin 2026-111-AWS, 11 September 2026.
12. Sizhe Chen, Yu-Lin Tsai, Ivan Evtimov, Kamalika Chaudhuri, Raluca Ada Popa, David Wagner, and Arman Zharmagambetov. 2026. *Repeat-After-Me: Black-Box Adaptive Visual Prompt Injection*. Meta AI research publication, 7 September 2026.
13. Jakub Pachocki. 2026. *An Alien Mind*. OpenAI, 6 September 2026. Accessed 14 September 2026.
14. Dario Amodei. 2026. *We Must Pace the Frontier*. September 2026. Accessed 14 September 2026.