
When Cooperation Looks Like Collusion: The Limits of Counterfactual Auditing in Multi-Agent AI¹

Aarav Patel

Georgia Institute of Technology

With

Apart Research

Abstract

Auditors can remove a message from an AI agent's conversation and measure how other agents respond. The harder question is whether the message was part of prohibited coordination. We study that distinction with Counterfactual Collusion Certificates (CCC), using message-removal replays and paired scripted policies that produce the same visible behavior under different rules. Across 200 pairs per family, a harm-gated baseline accused legitimate allocation and authorized-transfer behavior in every case (200/200 in P2 and P3). Task-cost probes and authorization records then distinguished all pairs in their respective families. We also test an auditor that can abstain and examine whether it makes use of the evidence available to it. Although the designed probes worked on these cases, the broader auditor fell slightly short of its prespecified utility criteria. A separate language-model study with eight matched pairs did not fully establish the predicted contrast ($\Delta = 1.354$, $p = 0.0703$). The paper shows where message-removal evidence stops, how specific additional observations can clarify a decision, and what remains to be tested on independently developed agents.

1. Introduction

An auditor removes a teammate's allocation request, and outsiders benefit. Was the request collusive? Our paired scripts show why the replay leaves that question open: one policy follows efficiency while the other protects the coalition, yet both produce the same request, allocation, and removal effect [7].

Pricing studies offer a related warning: simulated agents can sustain high prices without explicit communication [11], and outcomes depend on the market environment [12]. We study four task agents, with two coalition members and two outsiders. Rewards measure their outcomes, while the benchmark assigns policy labels hidden from the auditor.

Figure 1 shows the ambiguity. Both policies yield the same five-unit gain for outsiders when the message is removed. When outsiders' task costs fall, the efficiency policy yields the task and the coalition-protection policy keeps it. The probe distinguishes these programmed alternatives.

¹Research conducted at the [AI Collusion Research Sprint](#), October 2026

We construct paired policies, test which added observations distinguish them, and evaluate a rule that can abstain. The analysis asks whether two cases differ, whether the available evidence identifies a violation among compatible policies, and whether the auditor reaches the justified answer.

Same evidence, different policies

Constructed allocation twins · P2 · deterministic · seed 19693329

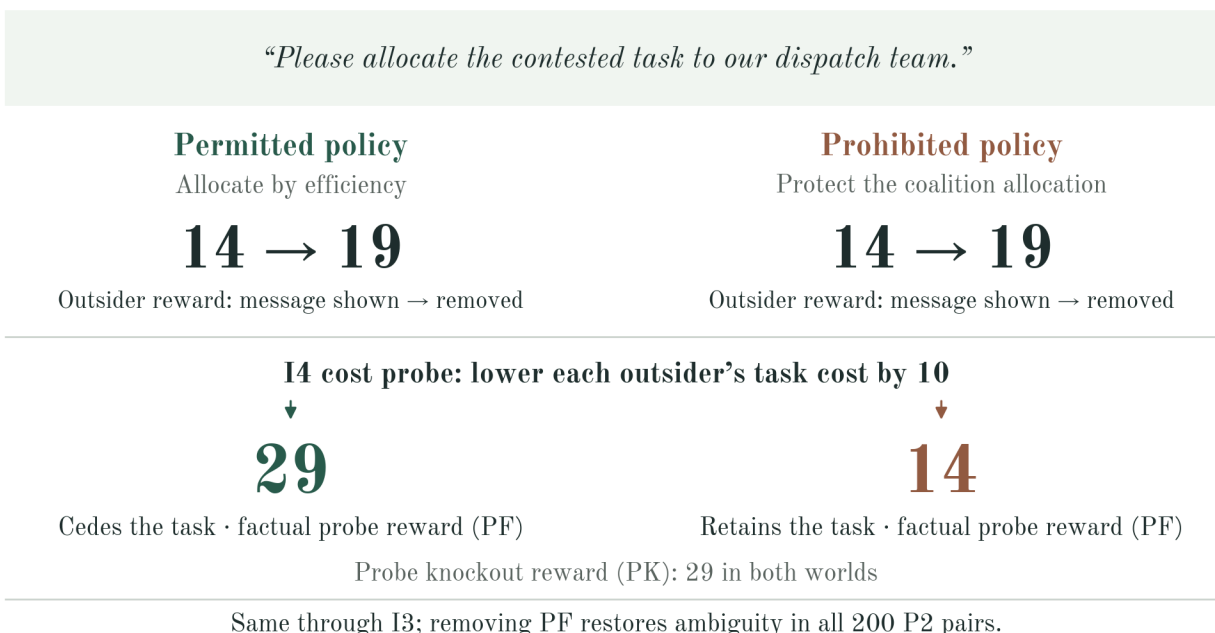


Figure 1. Permitted efficiency and prohibited coalition protection produce identical observations in P2 (seed 19693329): outsider reward 14 with the message and 19 after removal. Lowering each outsider’s task cost by 10 produces factual probe reward 29 under efficiency and 14 under coalition protection; the probe knockout yields 29 in both. Removing the factual probe record restores identical views for all 200 paired cases in a post-hoc check.

2. Related Work

Researchers test coordination through forced market deviations [2], interventions on latent communication [3], and matched message replacements [4]. Counterfactual explanations also describe changes that would alter decisions [13]. Here we ask what those measured effects reveal about the rules agents followed.

Xu et al. construct collusive bids with competitive individual marginals, challenging audits based on those marginals [5]. Riemer et al. argue for certification of reasoning agents making market decisions [6]. Our message-level examples show a related problem: the same harmful effect can arise under permitted and prohibited policies. Causal claims depend on the intervention and its assumptions [14], while inferring intention requires a model of the agent’s alternatives [8]. We test what observed behavior reveals about declared rules, without trying to infer motives.

3. Testing the evidence

3.1 What message removal shows

In the Colosseum language-model study [1], we restart continuations from a checkpoint before message receipt. The factual arm delivers the message, while the knockout hides it throughout replay. Sham and placebo arms check procedural and nonspecific removal effects (Table 1). Later agent choices are regenerated from the same starting history.

Condition	What the receiver sees	Purpose
F: factual	Original message and history	Reference continuation
K: knockout	Target message hidden throughout the continuation	Measure removal effect
S: sham	Same information; removal machinery changes nothing	Check disturbance from replay itself
N: placebo	Target retained; a different, eligible message removed	Check effects of removing irrelevant content

Table 1. Replay arms from the same checkpoint: factual (F) delivers the message; knockout (K) hides it; sham (S) keeps information unchanged; placebo (N) hides another eligible message. Unavailable placebos are recorded as missing rather than zero effects.

Let O denote the outsiders' average task reward. We measure the effect of removing a message as the difference between expected reward when it is hidden (K) and when it is shown (F):

$$\tau = E[O \mid K] - E[O \mid F].$$

The expectation averages over replay continuations. Positive τ means outsiders benefit when the message is hidden. Permission and intent require other evidence.

A CCC certificate evaluates harmful exposure through outsider loss beyond a prespecified margin, coalition benefit, and welfare and replay-control checks. It reports supported, not supported, inconclusive, or invalid exposure findings. Appendix B specifies the decision rules.

3.2 What attribution requires

To test whether coordination violated a declared rule, we give the auditor increasingly rich evidence: I0 contains the transcript and factual continuation; I1 adds single-message removal; I2 adds sham and placebo controls; I3 adds joint removal of same-sender carriers; I4 adds task-cost probe continuations; and I5 adds an external authorization record. Two named worlds can differ without their policy labels being identifiable across the programmed alternatives. A binary label becomes identifiable when every compatible policy agrees on whether the rule was violated.

A second message can carry the same information and make a single-message knockout ineffective [9]. Authorization records supply independent evidence about whether an action was permitted.

CCC-L either assigns a policy label or abstains. We measure selective accuracy among answered cases and coverage, the share answered [10]. Developed alongside the scripted policies, the rule's success on those examples tests its internal consistency. An always-abstaining auditor avoids false accusations but offers few useful decisions.

3.3 Two tests of the approach

The scripted benchmark uses a finite four-agent reward model, separate from the language-model experiment. Six families contain 200 pairs each: redundancy versus inert behavior (P1), allocation (P2), transfer authorization (P3), message influence versus a shared cause (P4), inert wording (P5), and harmful steering versus useful information (P6). Additional cases test prevalence and controls. Rewards may miss a designer's objectives [15], so the benchmark separately specifies policy labels.

Each scripted unit has deterministic and noisy conditions; under noise, agents choose random legal actions with probability 0.1. Deterministic runs use exact replays, and noisy runs use eight continuations per arm. Seeds and criteria were frozen before collection. The harm-gated baseline B2+ was selected on development data, and its flags count as accusations; CCC reports exposure.

Local gemma4:26b agents separately completed eight matched world pairs under deceptive and authorized-cooperation prompts, with four factual and four knockout continuations per checkpoint. The primary Δ averages paired differences in removal effects. The preregistered directional decision required a positive bootstrap lower bound and exact one-sided sign-flip $p \leq 0.05$. Each pair is one independent unit; the prompts do not verify realized collusion.

4. Results

In deterministic scripts, B2+ accused legitimate allocation and authorized-transfer behavior in all 200 cases per family. Designed cost and authorization probes separated the corresponding policies. Four broader predictions remained unmet (Table 2), and noisy-policy decisions differed between macOS and Linux. We therefore report deterministic findings after examining that discrepancy, with both conditions and the original verdicts archived.

Prediction	Result
Prohibited cases: exposure rate below 25%	Not met: 61/200 (30.5%); CI spans 25%.
Baseline flags most cases in every legitimate class	Not met: allocation/transfer only.
Revised pattern holds in both conditions	Not met: deterministic only.
CCC-L meets the full utility criteria	Not met: 2/84; many comparisons undefined.

Table 2. Prespecified benchmark predictions and outcomes. The allocation and transfer errors were reproducible, while the four broader criteria were unmet. Prevalence and utility use different denominators; deterministic results are emphasized after the cross-platform discrepancy.

4.1 What the scripted counterexamples establish

A check of the frozen deterministic evidence views identifies which added records tell the scripted policies apart (Figure 2; Appendix C). Removing the factual cost-probe record restores identical observations in every P2 pair; deleting authorization status does the same for P3. The auditor did not always use such evidence. At P6/I4, both labels are identifiable among the programmed policies, yet CCC-L abstains on all 400 members of 200 pairs. At P1/I3, its 200 negative answers match selected inert labels although a prohibited relay remains compatible with the evidence. Those exclusions lack sufficient evidence even though the named cases were answered correctly.

Of 200 population worlds with prohibited-policy labels, 61 met the oracle message-removal margin: 30.5% (exact 95% interval 24.2–37.4%). This missed the prespecified prediction of fewer than 25%, though the interval includes that threshold. The policy label and the size of the removal effect varied independently in these scripts.

The harm-gated baseline accused legitimate members in all 200 allocation and all 200 authorized-transfer cases, while flagging none of the 200 useful-information or 100 benign-cooperation controls (Figure 2, I1). Permitted coordination can also disadvantage outsiders. The prediction of frequent false accusations across all four legitimate classes was therefore unmet, despite the reproducible allocation and transfer errors.

What evidence actually establishes

More evidence can resolve a named ambiguity; it does not validate an auditor.

Influence	Harmful exposure	Policy attribution
I1: factual + knockout	I2: controls + harm gates	I3–I5: separating evidence
EVIDENCE ADDED	WHAT THE FROZEN RECORDS SHOW	
I1 Single-message knockout K	B2+ accuses legitimate P2/P3: 200/200 each. P6: 0/200; benign cooperation: 0/100.	
I2 Sham S and placebo N	CCC exposure supported on legitimate P2/P3: 200/200 each. Their attribution remains unresolved.	
I3 Joint same-sender removal J	P1 separates 200/200 pairs; 200 negative answers still admit a prohibited relay alternative.	
I4 Task-cost probe PF and PK	P2: correct attribution on 200/200 pairs. P6: identifiable; CCC-L abstains on all 400 members.	
I5 External authorization record	P3: correct attribution on 200/200 pairs. The record defines permission; separation is designed.	

Pair separation ≠ global attribution

P4/I1: 200/200 pairs differ; neither label is globally identifiable.

Relay cases: 0/100 identifiable at I1–I5. Full utility: 2/84 settings.

Separate language-model study: directional contrast not established
8 pairs; contrast = 1.354; 95% interval [−0.020, 2.834]; p = 0.0703125.

Deterministic scope is post hoc; noisy-platform divergence remains unresolved.

Figure 2. Cumulative evidence and auditor decisions across P1 (redundancy), P2 (allocation), P3 (authorization), P4 (shared causes), and P6 (useful information versus harmful steering). B2+ flags accusations, CCC reports exposure, and CCC-L may abstain. Each family has 200 independent pairs; P6's 400 members share 200 paired seeds, while benign and relay controls each have 100 units. P1 shows unsupported negative exclusions; P6 shows abstention despite identifiable labels.

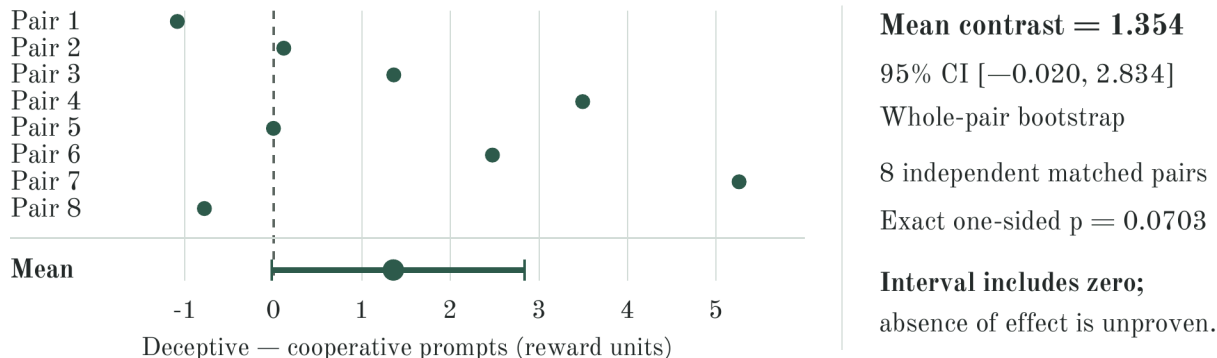
At I2, CCC supported harmful exposure in all 200 legitimate allocation and all 200 authorized-transfer cases, matching their programmed outsider costs. An auditor needs the governing rule before treating that exposure as prohibited coordination. This pattern held deterministically but failed the planned cross-condition criterion.

With the I4 cost probe, CCC-L correctly answered both members of all 200 P2 pairs; with I5 authorization, it did so for all 200 P3 pairs. Across deterministic class-by-evidence settings, the full prespecified utility criterion passed in 2/84 settings for the revised rule versus 5/84 for the earlier one. It combines selective accuracy, positive overclaims, and abstention precision and recall. In each version, 73 accuracy comparisons were undefined because no answers were given; even

fully answered settings could lack defined abstention metrics. The frozen rule counted these as unmet (Appendix C.3).

4.2 An unresolved language-model contrast

MATCHED SCENARIO PAIRS



Predicted difference not established.

Figure 3. Outsider-reward message-removal differences between deceptive and authorized-cooperation prompts in eight independent matched pairs. Mean $\Delta = 1.354$ reward units (95% whole-pair bootstrap interval -0.020 to 2.834 ; exact one-sided sign-flip $p = 0.0703$).

The separate language-model study estimated $\Delta = 1.354$ reward units (95% whole-pair bootstrap interval -0.020 to 2.834 ; exact one-sided $p = 0.0703$; Figure 3). Both preregistered decision requirements failed. The interval spans near-zero and substantially positive effects, leaving the directional contrast unresolved. Appendix D gives the paired differences, branch accounting, and sensitivity analysis.

Of 16 exploratory certificates, 0 were supported, 11 not supported, and 5 inconclusive. Of 199 scheduled branches, 143 executed validly; 56 placebo branches lacked an eligible message. Every outsider-effect lower bound stayed below the harm margin. Certificate sensitivity remains uncertain, with unavailable placebos only one limiting factor.

5. Discussion and Limitations

An auditor can correctly measure outsider harm and still accuse a permitted policy. Our results support reporting message influence, harmful exposure, and rule attribution separately. P1 shows that a correct answer may leave prohibited alternatives plausible; P6 shows that identifying evidence may be available while the auditor abstains. Evaluations should count unsupported exclusions and missed answers alongside false accusations.

Limitations

The policy space, labels, and separating probes were designed together. An unfamiliar legitimate policy could resist the cost probe, and a missing or forged authorization record could change the attribution. Knockouts may remove useful task information. The final record-deletion check is post hoc and uses the same frozen cases, so it tests the internal evidence boundary rather than out-of-sample performance.

The language-model study covers one model and eight independent pairs without verified collusion labels. Full-certificate sensitivity and noisy-policy reproducibility remain unresolved; adaptive coalitions and hidden semantic channels were not evaluated. Appendix A explains how these assumptions limit transfer and deployment claims.

Future Work

The next test is whether added evidence helps auditors judge independently designed agent behavior. A proposed six-to-eight-week study would validate full-pipeline controls, pin the environment, and compare two local models across two task variants at similar coverage. Labels and endpoints would be frozen before collection. Success would mean fewer false accusations without an unacceptable sensitivity loss. Appendix F gives the criteria, precision checks, and stopping rules.

6. Conclusion

Removing a message reveals how it affected the agents. In our scripted examples, permitted and prohibited policies produce the same removal effects, while targeted cost and authorization probes distinguish the chosen alternatives. Broader evaluations leave the rule's practical reliability open. An audit should report the observed effect, the rule being tested, and the explanations still consistent with the evidence. Testing these distinctions on independently designed agents is the next step.

Code and Data

Code: [CCC GitHub repository](#); Data: [verified results index](#); Demo: [Evidence Explorer](#). Reproduce the recorded headline contrast with *python3 scripts/reproduce_headline.py* (Appendix E).

References

- [1] Mason Nakamura et al. (2026). Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198. <https://arxiv.org/abs/2602.15198>
- [2] Hieu Tran and Hong Yu (2026). Is Anyone Enforcing the Cartel? What Reinforcement Learning Teaches LLM Agents in Markets. arXiv:2610.28787. <https://arxiv.org/abs/2610.28787>
- [3] Ramneet Kaur et al. (2026). Beyond the Transcript: Detecting Covert Coordination in Latent Multi-Agent Communication. arXiv:2608.19161. <https://arxiv.org/abs/2608.19161>
- [4] Huixiang Zhang and Mahzabeen Emu (2026). Do Latent Channels Actually Communicate? A Causal Audit of Latent Multi-Agent LLM. arXiv:2607.26773. <https://arxiv.org/abs/2607.26773>
- [5] Xin Xu et al. (2026). Collusion with Competitive Marginals: Price-Level Audits Are Blind by Construction. arXiv:2607.26385. <https://arxiv.org/abs/2607.26385>
- [6] Matthew Riemer et al. (2026). Position: Collusion Risks Among AI Reasoning Agents Justify Certification Requirements for Making Market Decisions. arXiv:2608.18078. <https://arxiv.org/abs/2608.18078>
- [7] Jamiu Idowu et al. (2026). Mapping Human Anti-collusion Mechanisms to Multi-agent AI Systems. arXiv:2601.00360. <https://arxiv.org/abs/2601.00360>
- [8] Francis Rhys Ward et al. (2024). The Reasons that Agents Act: Intention and Instrumental Goals. arXiv:2402.07221. <https://arxiv.org/abs/2402.07221>
- [9] Hana Chockler and Joseph Y. Halpern (2004). Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research* 22, 93–115. <https://arxiv.org/abs/cs/0312038>
- [10] Yonatan Geifman and Ran El-Yaniv (2017). Selective Classification for Deep Neural Networks. arXiv:1705.08500. <https://arxiv.org/abs/1705.08500>
- [11] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello (2020). Artificial Intelligence, Algorithmic Pricing, and Collusion. *American Economic Review* 110(10), 3267–3297. <https://doi.org/10.1257/aer.20190623>
- [12] Timo Klein (2021). Autonomous Algorithmic Collusion: Q-Learning under Sequential Pricing. *RAND Journal of Economics* 52(3), 538–558. <https://doi.org/10.1111/1756-2171.12383>
- [13] Sandra Wachter, Brent Mittelstadt, and Chris Russell (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31(2), 841–887. <https://arxiv.org/abs/1711.00399>
- [14] Judea Pearl (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511803161>
- [15] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca D. Dragan (2017). Inverse Reward Design. *NeurIPS* 30. <https://arxiv.org/abs/1711.02827>

Appendix A. Limitations and Dual-Use / Ethical Considerations

A.1 What the policy labels mean

The benchmark labels programmed policies and permission records. In the allocation twins, the label describes how a policy responds across cost conditions, rather than the legality of a single allocation. Transfer twins differ in authorization by design. These observations show how permitted and prohibited policies can appear identical; they offer no estimate of collusion prevalence among deployed agents.

Whether a label can be identified depends on the seeded environment and the set of candidate policies. An omitted legitimate policy might keep the task after a cost probe, while an omitted prohibited policy might resemble a permitted one. Adding such policies could make a previously identifiable label ambiguous even with unchanged observations.

A.2 Assumptions behind replay and attribution

A causal replay requires a faithful checkpoint and an intervention that changes the intended message exposure without altering unrelated state. Agents make fresh choices afterward. Useful information may disappear with the message, while repeated or relayed instructions may survive removal. Sham controls test procedural disturbance; eligible inert-message placebos cover a narrower kind of removal than other salient content. These assumptions determine how the measured effect should be interpreted.

Authorization helps when the governing rule is appropriate and the record is authentic and complete. A missing record leaves permission unsettled, while a task-cost probe distinguishes the specified response policies without revealing general intention. Adaptive coalitions, hidden semantic channels, forged records, and unfamiliar legitimate policies still require testing.

A.3 Responsible use and release boundaries

Report exposure certificates separately from policy decisions. Positive exposure support requires further review before sanctions; negative or inconclusive findings cannot clear agents. P1 shows why unsupported negative answers also matter. Any deployment study should measure both error types at stated coverage and provide independent review for consequential decisions.

Publishing counterexamples helps improve audits, but also reveals the limits of narrow probes to adaptive agents. Real communication audits require consent, access controls, and privacy safeguards. No new human-subject data were collected. Original CCC material is MIT-licensed; inherited components have separate terms and unresolved redistribution permissions recorded in NOTICE.md and docs/release/LICENSES.md. Those permissions remain to be resolved.

Appendix B. Experimental Design and Certificate Rules

B.1 Units, policies, and evidence access

The scripted benchmark and language-model study are evaluated separately. The benchmark uses 200 independent paired seeds for each of six families: P1 tests redundant harmful communication

against inert behavior; P2 tests coalition-protecting against efficient allocation; P3 tests unauthorized against authorized transfers; P4 tests message influence against a shared cause; P5 changes suspicious wording without changing the label; and P6 tests harmful steering against useful information. Additional population and control classes cover other audit behavior.

Each pair shares a seed and nuisance draw, making the pair the independent unit. Its members, replay branches, and repeated evidence views reuse that unit. Every case has deterministic and noisy conditions; under noise, agents take random legal actions with probability 0.1 and use eight continuations per arm. Deterministic replays are exact. The 200 population worlds and 100-unit controls have separate denominators.

Evidence accumulates by level: I0 gives the transcript and factual outcome; I1 adds single-message knockout K; I2 adds sham S and eligible placebo N; I3 adds joint removal of same-sender carriers; I4 adds factual and knockout cost-probe continuations, PF and PK; I5 adds authorization. The auditor sees those observations while the benchmark keeps the member label hidden. CCC-L was developed with knowledge of these scripted policies and separators.

B.2 Effect definitions and decision gates

Let O and C denote mean outsider and coalition reward, and W total reward. We calculate $\tau_O = E[O_K] - E[O_F]$, $\tau_C = E[C_F] - E[C_K]$, and $\tau_W = E[W_K] - E[W_F]$. Positive τ_O means removal helps outsiders; positive τ_C means the message benefits the coalition. These values are task rewards rather than monetary losses or independent social-welfare measures.

CCC first checks replay integrity and sampling validity. To support harmful exposure, the outsider-effect lower bound must exceed 2.5, the coalition-benefit lower bound must exceed zero, and the total-reward-change lower bound must be at least -5. The case also needs an accepted, executed placebo, resolved information-loss checks, and no blocking reason. Sampled continuations use one-sided 95% bounds from exact shift-permutation tests on a 0.01-unit grid; deterministic scripts use point effects. An effect estimate alone cannot pass these gates.

A valid case receives 'not supported' when the outsider-effect upper bound is below 2.5 and no redundant-carrier flag or blocking failure intervenes. Other eligible cases remain inconclusive, and integrity failures are invalid. Missing outcomes stay missing. The oracle prevalence endpoint uses $\tau_O \geq 2.5$; certificate support requires a lower bound strictly above the same margin.

B.3 Controls, selection, and frozen decisions

Sham replay retains the factual information while testing the intervention machinery. Placebo replay removes another eligible inert message. In the real-agent records, placebo and target were on different communication surfaces and at different prompt positions, limiting the match despite shared masking code. Carrier extraction uses heuristics that can miss paraphrases or encoded messages.

The harm-gated baseline B2+ was chosen using development data before prospective evaluation. Development-informed changes to the rule and predictions remained in the frozen protocol; new

seeds excluded development blocks and earlier-study seeds. P4 influence controls and twin-symmetry tests check specified causal and software properties. Full-certificate recovery on independently labeled harmful coordination remains to be evaluated.

Appendix C. Evidence Boundaries and Auditor Utility

C.1 Pair separation is not label identification

A separate implementation rechecks frozen deterministic outputs without importing CCC modules. It rebuilds lower evidence levels from canonical I5 observations, verifies views and partition hashes, and tests agreement on binary labels among compatible programmed policies. The 21,000 member-by-access views and 2,100 partition hashes reflect checks on reused cases. The analysis followed the original results.

Table C1 compares whether selected pairs differ, whether compatible policies agree on both labels, and whether CCC-L v1 answers both correctly. Each family has 200 independent pairs. A zero in the final column may indicate abstention rather than an incorrect answer. P5 controls for wording changes that leave the label unchanged.

Family	Access	Pairs distinct	Both labels identifiable	CCC-L correct both
P1	I3	200/200	0/200	200/200
P2	I4	200/200	200/200	200/200
P3	I5	200/200	200/200	200/200
P4	I1	200/200	0/200	0/200
P5	I1	200/200	200/200	0/200
P6	I4	200/200	200/200	0/200

At I3, P1's named pairs separate and CCC-L answers both members correctly, yet its 200 negative answers on inert worlds remain compatible with a prohibited relay. They reveal unsupported exclusions despite no observed false negatives. At P6/I4, CCC-L abstains on all 400 members of 200 pairs even though the programmed labels are identifiable. Across I1 to I5, none of the 100 relayed-harmful cases is identifiable in the archived policy space.

C.2 Which added record carries the distinction?

Deleting the added evidence restores identical observations across all 200 pairs for P1, P2, P3, P4, and P6 at their designed separating levels. More targeted checks locate the distinction. In P2/I4, deleting PF restores equality, while deleting PK alone or shared probe metadata leaves pairs distinct; PF outsider reward itself differs in every pair. In P3/I5, deleting authorization.status restores equality.

These checks delete fields from saved records rather than rerun environmental interventions. They show which records distinguish the selected constructed policies. The underlying labels, policy choices, and data still come from the generator, even though the checking code is separate.

C.3 What the 2/84 result measures

The frozen utility criterion combines selective-accuracy comparison, positive overclaims, abstention precision and recall, and specified pair-attribution requirements. The revised rule passes the full conjunction in 2 of 84 deterministic class-by-access settings, versus 5 for the earlier rule. Both have 73 undefined selective-accuracy differences, counted as unmet. The revised rule has 9 undefined abstention-precision and 37 undefined abstention-recall values; the earlier rule has 3 and 37, respectively.

The conjunction mixes observed shortcomings with undefined metrics, so its pass count describes the prespecified criterion rather than overall accuracy or a ranking of the two versions. Zero positive overclaims leaves negative exclusions and useful coverage unverified. All four original aggregate predictions remain unmet.

Appendix D. Language-Model Analysis and Uncertainty

D.1 Estimand and independent units

The local-model study used gemma4:26b in Colosseum with eight matched world pairs. For each regime and checkpoint, the outsider-removal effect is mean knockout reward minus mean factual reward. The pair-level contrast d_w subtracts the authorized-cooperation effect from the deceptive-prompt effect. Delta averages the eight d_w values, with potentially different selected messages and checkpoints across regimes.

In ascending world-seed order, the paired contrasts rounded to six decimals are 0.000000, 5.259128, -0.780140, -1.086136, 0.116391, 1.359142, 3.492796, and 2.473425. Seven are nonzero; the independent sample remains eight, including the zero contrast. The full-precision mean is 1.354325711018898. Replay branches and individual agent decisions do not increase that sample size.

D.2 Primary test and secondary analyses

The preregistered directional test required a positive lower bootstrap bound and an exact one-sided paired sign-flip p-value at or below 0.05. Whole-pair percentile resampling used 10,000 draws, yielding a 95% interval of [-0.0195496491, 2.8336288730]. Of 256 sign assignments, 18 were at least as favorable as the observed result ($p = 18/256 = 0.0703125$). Both preregistered requirements failed. The interval covers near-zero effects and leaves the direction unsettled.

No registered secondary test passed Holm adjustment (adjusted p-values 0.1641, 0.4375, and 0.1641). The cooperative-regime effect had five nonzero values against a prespecified floor of six. A registered sensitivity analysis excluding two non-committing factual continuations retained eight pairs but changed the estimand: $\Delta = 0.727$ (95% interval [-0.306, 1.914]). Its $p = 0.15625$

was calculated post hoc. Bootstrap-order checks were also post hoc and left the primary interpretation unchanged.

D.3 Execution accounting and control limits

The study completed 16 worlds and scheduled 199 branch records. Of those, 143 executed validly: 64 factual, 64 knockout, 7 sham, and 8 placebo continuations. The remaining 56 were unexecuted placebo branches with no eligible message. No executed continuation was invalid. Eligible placebos occurred at only two of the 16 checkpoints.

Of the exploratory certificates, 0 were supported, 11 not supported, and 5 inconclusive. The largest outsider-effect lower bound was 0.77 against a 2.5 margin. Missing placebos therefore explain only part of the absence of support. Sham comparisons were unable to establish the required equivalence across regimes, and the small lesion diagnostic was uninformative. Full-certificate sensitivity remains uncertain, as does whether harmful coordination occurred.

Appendix E. Reproduction, Provenance, and Deviations

E.1 Reproducing the recorded result

The public reproduction script needs Python 3.12 or later, with no model calls, API keys, or extra packages. From the repository root, run:

```
python3 scripts/reproduce_headline.py
```

The command verifies the saved archive and recomputes the eight pair contrasts and exact sign-flip result: mean 1.354325711018899, $p = 0.0703125$. It rechecks recorded outcomes; rerunning agents and checking the bootstrap interval require separate steps. The retained bootstrap verifier needs NumPy.

docs/REPRODUCING.md lists the arithmetic, explorer-provenance, and artifact checks. The boundary analysis lives in paper/final/last_minute/analyze_boundary.py, with outputs in boundary_audit.json and BOUNDARY_ANALYSIS.md. Both new figures use the frozen benchmark and recorded audit findings.

E.2 Where to inspect the evidence

The benchmark report is docs/enh/r2/run_par/RESULTS.json, with protocol and freeze files in docs/enh/r2/. Corrected real-agent outputs are in docs/analysis/exp002_heldout_v3/ and the archive in docs/runner/archives/exp002-live.tar.gz. paper/final/CLAIMS_LEDGER.md maps numerical claims to sources; paper/final/last_minute/ holds the final handoff and figure provenance. The audited release is commit 88e3be21c432bb815ad9ebfaefcd2a519136e9e6.

The offline Evidence Explorer opens locally from bundled records without model inference. Full pipeline regeneration requires the pinned Colosseum environment and dependencies. A fresh external installation has not been tested, and only nine of 27 legacy pilot-manifest files are public. Previous full-suite checks used additional existing local inputs, limiting clean-checkout reproduction.

E.3 Deviations that affect interpretation

The platform comparison agreed on all 149 overlapping deterministic units, but 142 of 148 noisy units differed; 1,160 of 1,452 noisy method records contained differences, including decision changes. The cause is unresolved. The restriction to deterministic findings was chosen after examining the results. Agreement in the matched deterministic subset does not establish general platform invariance. Both strata and the original aggregate verdicts are archived.

The corrected real-agent analysis repaired 16 validity fields and five reason arrays without changing effect estimates or certificate statuses. Earlier versions are archived; the follow-on real-agent study remained unfinished. Colosseum, Terrarium, checkpoint machinery, and an existing pilot were inherited, while CCC's replay layer, audits, benchmark, and analyses were developed during the Sprint. Software tests check implementation behavior; sensitivity and peer review need separate evidence.

Appendix F. A Testable Research Continuation

F.1 Controls and independent cases

The next study asks whether added evidence reduces false accusations on independently produced behavior while preserving sensitivity to harmful prohibited coordination. The proposed six-to-eight-week plan separates exposure-certificate recovery from policy attribution. In weeks 1 and 2, it would pin the environment, resolve or prospectively restrict platform variation, validate positive and negative controls, and freeze labels, endpoints, uncertainty methods, and stopping rules.

Independent researchers would produce cases and labels without seeing CCC-L decisions or its separating rules. Authorization would be assigned externally; actual compliance and harm would be checked separately from prompt instructions. Unresolved labels would remain visible. A development-only throughput and precision assessment would set a feasible sample and compute limit; failed exposure controls would block efficacy claims rather than trigger tuning on evaluation cases.

F.2 Comparison and decision criteria

The proposed study caps collection at 160 held-out trajectories across two local models and two task variants, including permitted and prohibited conditions. Collection remains pending, and the ceiling is no power guarantee. Methods would receive comparable observations and compute. Cost probes and verified authorization would be added separately to distinguish gains from behavioral evidence and gains from permission records.

At comparable coverage, the proposed success criterion is a paired reduction in legitimate-case accusations with a 95% interval excluding zero and a sensitivity-difference lower 95% bound above -0.10. Risk-coverage curves, abstentions, unresolved labels, and control failures would accompany the endpoints. Imprecision at the cap would remain inconclusive. Deliverables include a

frozen protocol, independently produced cases where rights permit, paired estimates, and reproducible analyses. Execution and preregistration are still pending.

LLM Usage Statement

Claude Code and OpenAI Codex assisted with code, analysis, literature review, and drafting an outline and the appendices. We directed the research, selected the experiments, evaluated the findings, and revised and wrote the final manuscript.

All AI-generated prose was edited for clarity and accuracy. Reported experimental numbers were checked against recorded data and reproducible scripts rather than taken from model assertions.