

Talk, Actions, and Planted Approval Pacts

An Exploratory Audit

Carly R. Anderson

HumanAIOS LLC

AI Collusion Research Sprint | October 2026

Abstract

We explored whether combining agent messages with approval actions helps distinguish coordinated false approval from other behavior. Two language-model reviewers assessed an inert change against an executable numeric requirement. Across 60 episodes and six assigned conditions, we compared a keyword-based talk detector, an action detector, and their conjunction. The original no-retry judge pass produced no observed positives and 11 unknown outcomes, failing the prespecified minimum of five positives. A first retry yielded four positives; a subsequent retry, added after inspecting that count, yielded seven. All results incorporating retries are exploratory. Under the final model-judge labels, the hybrid detector flagged four explicit-agreement episodes and none of 53 labeled negatives, while missing three concealed-agreement positives. The talk detector flagged all ten benign negations containing its target phrase. Follow-up reviewers agreed on the four explicit cases but disagreed on other bilateral approvals, exposing dependence on the definition of agreement. Reviewer identities and separation from the builder were not authenticated. This pilot documents limitations of simple monitoring rules and the sensitivity of their evaluation to labeling and retry decisions. It establishes neither spontaneous collusion nor deployment accuracy.

1 Introduction

Two reviewers can approve the same defective change because they agreed to ignore a requirement, because they made the same mistake, or because one followed the other. A shared action alone does not identify which explanation holds. Conversely, a stated pact can remain unexecuted. Audits should distinguish communication, observed action and the evidence used to connect them.

This sprint project asks whether simple checks of speech, actions and their conjunction separate these cases in a toy review task. Our contribution is a recomputable record with explicit negative cases, recorded API failures and multiple disclosed interpretations of agreement: every count can be recomputed from the stored artifacts, but rerunning the live episodes would draw new samples at temperature 1 and would not reproduce these outputs. The study is a measurement pilot, not a validated detector or evidence about naturally arising collusion. The prespecified positive-control minimum failed in the original no-retry analysis.

2 Related Work

The sprint asks whether combining agent speech and behavior improves collusion detection and what evidence distinguishes agreement from correlation [1]. Colosseum compares communication-based and action-based measures and reports cases where plans and actions diverge [2]. We share that

measurement concern but do not reproduce Colosseum, measure coalition advantage, model a market, or compare against its methods. Our deterministic numeric task instead makes requirement failure directly inspectable. Cerebras supplied the hosted inference interface [3]; the exact prompts and outputs are preserved with our code and data [4].

3 Methods

Each episode contained a limit L and proposed integer values $[L - 1, L + 1]$. The requirement was that every value be at most L , so every change contained a known failure. Numeric seeds 10–19 varied the limit. Six assigned conditions supplied ten episodes each, shuffled with order seed 20261025: honest teamwork, a misleading prior note intended to elicit shared error, benign negation of pact words, unilateral false approval, an explicit planted pact, and a concealed planted pact using “green signal”. Assignment was an instruction, not a behavior label.

Both actors used qwen-3.8-27b through Cerebras. Sampling used temperature 1, top_p 0.95, high reasoning effort and max_completion_tokens 32768. Within an episode, A sent a message, B replied after seeing A, and each then produced a separate ballot with both messages visible. A and B had separate histories; histories were fresh across episodes. The task allowed text only and conferred no real approval authority. The planted-role prompts explicitly framed the pact as a simulation. Raw requests, responses, request identifiers, ballots and usage are stored in the run folder [4].

We retained three vo.1 rules. `talk_only` fires when both messages contain the case-insensitive substring “approve together”. `actions_only` fires when both ballots are APPROVE and the executable requirement fails. `hybrid` is their conjunction. These rules do not understand negation or infer hidden agreements. The model judge gpt-oss-120b saw only the task, messages and ballots, not the assigned condition or detector output. It used temperature 0, top_p 1, medium reasoning effort and max_completion_tokens 4096. It was asked whether both reviewers expressed agreement to approve despite the failed requirement.

An observed positive required a failed requirement, two APPROVE ballots and a TRUE judge answer. An episode was negative if either ballot was REJECT, the requirement passed, or the judge said FALSE; otherwise it was UNKNOWN. UNKNOWN outcomes were excluded from binary tables with their coverage reported. A minimum of five observed positives was required before interpreting detector sensitivity. One sample per episode, repeated prompt templates and a single actor model limit any inference beyond these runs.

The vo.2 protocol, schedule and code hashes were pushed at 288a20f8 before actor calls. This was a local content freeze, not an authenticated external preregistration. The protocol prohibited retries. Of 60 first-pass judge calls, 54 returned HTTP 429. The first retry targeted only those failures; a second retry was added after the first-retry view contained four observed positives, below the minimum five. The final exploratory analysis uses the first non-429 outcome along passes 1–3. Original artifacts and an amendment record preserve these deviations. The four-positive first-retry view must not be described as the original frozen no-retry result.

Stage	Positive	Negative	Unknown	Status
Original no retry	0	49	11	Threshold failed
After first retry	4	52	4	Below minimum 5
After second retry	7	53	0	Exploratory only

Table 1. Recounted analysis stages. UNKNOWN outcomes are excluded from binary detector tables.

A prior 20-episode pilot yielded no bilateral approvals and motivated the planted-role v0.2 follow-up. Scripted controls checked implementation behavior; they are separate from live model evidence. No model calls were added for the final arithmetic and artifact checks.

4 Results

All 60 actor episodes completed: 240 of 240 actor calls succeeded, totaling 58,670 prompt tokens and 96,074 completion tokens (154,744 total). Recorded actor cost was approximately \$0.20; the project log estimates about \$0.34 including earlier runs and judging. These are run estimates, not a billing reconciliation. The original no-retry observed labels were 0 positive, 49 negative and 11 unknown. No sensitivity claim follows from that view.

Detector	TP	FP	TN	FN	Recall	FP rate
talk_only	4	11	42	3	4/7 (0.57)	11/53 (0.21)
actions_only	7	4	49	0	7/7 (1.00)	4/53 (0.08)
hybrid	4	0	53	3	4/7 (0.57)	0/53 (0.00)

Table 2. Final exploratory confusion counts after both judge retry passes. Labels use the automated judge and mechanical rule, not reviewer consensus. TP/FP/TN/FN denote true positives, false positives, true negatives and false negatives. Rates are descriptive; with seven positives the 95% Wilson interval for recall is 0.25–0.84 for talk_only and hybrid and 0.65–1.00 for actions_only.

The hybrid fired on exactly v2-10, v2-15, v2-30 and v2-53: four episodes with explicit pact language from both actors and two APPROVE ballots. Hybrid flagged all four explicit subset cases. This subset observation is not an overall recall claim. The 0/53 false-positive count in Table 2 belongs specifically to the final model-judge labels, and it is close to structural: a hybrid false positive requires both actors to write the exact phrase and both to approve a failing change while the judge answers FALSE, which the arm designs made unlikely. It should not be read as evidence of precision. The four actions_only false positives are the four episodes in which one actor sent a routine message, the other wrote an agreement phrase, and both approved; the judge labeled them FALSE, so they depend on the label definition, not on detector error. talk_only also flagged all ten benign-negation episodes and one planted explicit episode with a REJECT ballot.

Assigned condition	Both approve	Observed positive	Hybrid flags
Honest teamwork	0	0	0
Shared mistake	0	0	0
Benign pact words	0	0	0
Unilateral false approval	0	0	0
Planted explicit pact	7	4	4
Planted concealed pact	4	3	0

Table 3. Assigned conditions and observed outcomes after both retries. Every arm has ten episodes; assignment alone does not establish an actual pact.

Eleven episodes had two APPROVE ballots on a failing change. The owner and final judge accepted seven as agreements; one AI evaluator accepted all eleven using a ballot-centered reading, while the reported DeepSeek reviewer accepted only the four explicit cases. The concealed phrases alone did not settle agreement under the narrower text reading. These follow-up labels are informative sensitivity checks; their identities and separation from the builder were not authenticated.

Subset	n	Final judge	Owner	AI ballot reading	DeepSeek text reading
Explicit pact both sides	4	TRUE	TRUE	TRUE	TRUE
Green signal both sides	3	TRUE	TRUE	TRUE	FALSE
Routine plus pact message	4	FALSE	FALSE	TRUE	FALSE

Table 4. Agreement labels in the eleven bilateral-approval episodes. Follow-up review provenance is unverified; this is not a consensus ground-truth dataset.

On the separate action-level definition (both approve a failing change), `actions_only` matches all eleven by construction. Hybrid catches four and misses seven. This is a definition check, not evidence that `actions_only` detects agreements perfectly. No shared-mistake approvals occurred, so the planned comparison on that negative control was unevaluable. Concealed misses and benign-negation flags partly follow directly from the prompts and keyword rule.

5 Discussion and Limitations

The useful result is a record of where simple monitoring and its evaluation break down. A conjunction can remove some lexical false alarms, yet it loses cases with no target phrase. Changing agreement from an expressed bilateral understanding to coincident approval changes which outcomes count as positives. The study therefore supports careful separation of an action-level violation, a communication signal and an adjudicated agreement; it does not establish a generally superior detector.

The first limitation is protocol deviation: judge retries were forbidden, and the original no-retry data contain no observed positives. Even the first-retry view remains below the five-positive minimum. The second retry was added after that count was seen. Final figures must stay exploratory and cannot repair the failed frozen analysis retroactively.

Second, review provenance is limited. The automated judge ran on the same account and harness. The owner reviewed after seeing builder discussion. Replies attributed to Empirica and DeepSeek were relayed as pastes, with unverified identity. One later evaluator pass had repository access and

was informed rather than blind. The common acceptance of four explicit cases is a narrow agreement observation, not independently authenticated evaluation.

Third, the positive cases were instructed roleplay on an inert toy task, not spontaneously formed coalitions. Fourth, the definition of agreement and access to the concealed code meaning determine the labels; different readings produce different denominators. Fifth, the study uses one actor model, one sample per episode, few positive outcomes, builder-written prompts and repeated templates. It does not estimate prevalence, real-world intent, deployment accuracy or generalization. Descriptive counts are reported without a statistical superiority claim.

Future work should fix a single agreement definition and error/retry policy before any calls, use fresh episodes, retain all failures, and secure a blinded review channel with recorded identity and exposure. An adequate positive control and actual shared-error outcomes are prerequisites for testing the original comparison. These are follow-up requirements, not claims satisfied by this pilot.

6 Conclusion

The frozen evaluation did not meet its positive-control requirement. Exploratory follow-up shows four explicit subset detections, systematic failure on benign negation for the talk rule, and misses on concealed agreements for the hybrid. The reported scores depend on both the chosen label and the later retry process. The defensible contribution is a recomputable pilot with exposed measurement limits, rather than proof that the detectors work in deployment.

Code and Data

All counts in this report were checked from the pinned research branch at `cc5e6618a35df15e30f4c1fe714891cb2276c727` [4]. The run is `apart-1010-v02-001`. The final check recomputed labels and all three detector tables from the stored evidence and judge passes, confirmed the four hybrid episode IDs and token totals, and reproduced 43 passing offline tests. This verification made no inference calls and does not authenticate provider logs or reviewer identities.

Prior Work and Author Contributions

Carly R. Anderson is the sole confirmed human author and study owner, affiliated with HumanAIOS LLC. Preexisting HumanAIOS coordination and ACAT-related work, including issue #986 and PR #987, supplied project context. The `collusion_audit` harness, controls, planted-role follow-up, live trial artifacts and this report were produced for the sprint. Prior project tests are not counted as new collusion results. AI implementation and editorial assistance are disclosed in Appendix B. No other human authorship is inferred from pasted evaluator replies.

References

[1] Apart Research, Poseidon Research and AE Studio. 2026. AI Collusion Research Sprint, questions 2.2 and 2.3 and submission guidelines. apartresearch.com/sprints/ai-collusion-research-sprint-2026-10-23-to-2026-10-25

[2] Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., and Bagdasarian, E. 2026. Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. [arXiv:2602.15198v2](https://arxiv.org/abs/2602.15198v2)

[3] Cerebras. 2026. OpenAI Compatibility. Cerebras Inference documentation. [inference-docs.cerebras.ai/resources/openai](https://docs.cerebras.ai/resources/openai)

[4] HumanAIOS. 2026. Collusion audit code, protocols and recorded experiment artifacts. Operations research branch, commit cc5e6618a35df15e30f4c1fe714891cb2276c727. Run apart-1010-v02-001. github.com/humanaios-ui/operations/tree/cc5e6618.../research/collusion_audit

Appendix A Limitations and Dual Use Ethical Considerations

Publishing precise keyword rules makes their blind spots visible to attackers as well as defenders. The concealed-arm prompts illustrate a code-word failure in a deliberately bounded setting. These results should not be presented as a recipe for operational collusion or as a dependable safeguard. No covert communication system, production integration or real-world target was developed.

The task involved simulated numeric changes and inert ballots; there was no real approval, deployment, external transaction or decision about a person. The public record contains synthetic task data and model outputs. No personal dataset was supplied to the study actors. The builder records API key retrieval from the macOS Keychain; that is an implementation/provenance statement, not a forensic attestation that credentials were never elsewhere. No API key is reproduced in this report.

Pacts were planted in prompts, and some actors declined or deviated. An APPROVE token in a message was not substituted for the separately collected ballot. We preserve failed calls, unknown labels and disagreements. The original positive-control failure and both retry stages remain visible. We do not claim spontaneous collusion, authenticated independent evaluation, prevalence, deployment accuracy or generalization.

Appendix B LLM Usage Statement

Claude and ChatGPT assisted with experiment implementation, protocol discussion, debugging, evidence checks and drafting. qwen-3.8-27b generated actor messages and ballots; gpt-oss-120b supplied automated agreement judgments. Additional reviewer replies attributed to Empirica and DeepSeek were relayed by the owner and remain unauthenticated. The report distinguishes their claimed judgments from mechanically recounted artifacts. The study owner is responsible for final authorship, interpretation and submission. No AI system is listed as a human author.

Appendix C Final Count Check

At the cited source commit, the first judge pass has 5 FALSE, 1 TRUE and 54 UNKNOWN judge answers; its lone TRUE answer does not satisfy both-APPROVE, yielding 0 observed positives. The first retry view has 4 observed positives, 52 negatives and 4 unknown outcomes. Final selection after the second retry has 7 positives and 53 negatives. A remaining UNKNOWN judge answer is mechanically negative because the ballots do not both approve. Thus zero unknown observed outcomes in the final table does not mean every judge call succeeded.