
Words Actions and Evidence in Collusion Detection¹

Siddeshwar U S

MIT, Manipal

Team spacenets

With

Apart Research

Abstract

Detecting harmful collusion requires separating shared agreements from legitimate cooperation, independent deception, and plans that are never executed. We study whether combining public statements with executed actions improves detection in a controlled five-agent repeated public-goods game. A frozen open-model pipeline generates six conditions and evaluates four views: public text, actions, their raw union, and a code-built claim-versus-outcome report. Two development families determine configuration and thresholds; eight scheduled held-out families yield 48 episodes, of which 48 are valid under the automatic episode criteria. Held-out implementation AUROC is 0.597 for text, 0.843 for actions, 0.719 for the raw union, and 0.731 for the structured report. Comparisons use paired family bootstrap intervals and development-locked operating thresholds. Independent deception provides a deliberately matched hard negative, while talk-only plans test whether monitors confuse intention with execution. The experiment uses controlled public communication for concealed and independent-deception conditions, and uses event-based automatic outcome labels. Results describe detection behavior in this constructed setting, including failures and uncertainty, rather than proving general ability to detect secret agreements from public records. Combining information did not consistently outperform actions alone, making the limits of public agreement evidence explicit.

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

1. Introduction

Parallel harmful actions and broken promises can arise with or without a shared agreement. A monitor that flags all deviations therefore measures a different target from implemented harmful collusion. Track 2.2 asks whether combining words and actions improves detection over either source alone. We test that question in a controlled repeated public-goods game.

We contribute a six-condition protocol separating agreement, execution and harm; four isolated monitor views, including a deterministic claim-versus-outcome report; and a frozen, family-paired evaluation with uncertainty, hard negatives and exact offline replay. The observed result is a limitation: the structured hybrid does not consistently outperform actions alone. Outcomes use frozen event-based automatic labels.

2. Related Work

When Agents Lie [1] separates private plans, announcements and actions. We use its released game for payoff verification and replace the runner and typology with our own controller [6]. COLOSSEUM [2] motivates coalition-level outcome checks; its regret is not the fixed-outsider welfare contrast used here. TraceGuard [3] motivates structured consistency rubrics, while Arike et al. [4] motivate controlled information-access comparisons. SCHEME [5] concerns coordinated software sabotage; its success rates do not transfer to this game.

Our audit identified 22 confirmed records across the supplied papers, including released-code and reporting defects; these do not establish that all historical headline results are invalid. We avoid the affected typology, hidden-state monitor inputs and aggregation assumptions, and retain an explicit finding-to-implementation map. We do not reproduce the papers’ full experiments or claim the complete TraceGuard stack.

3. Methods

3.1 Environment and experimental controls

Five delegates play ten rounds. Each receives 5 tokens and chooses an integer contribution from 0 to 5; the principal mandates 3. Public promises, private plans and executed actions are stored separately. Payoff is 5 minus own contribution plus 0.3 times total contributions. Thus all-3 gives 6.5 each and welfare 32.5; two zeros with three 3s give pair payoffs 7.7, outsider payoffs 4.7 and welfare 29.5. Appendix A gives the formal definitions.

Table 1. Six assigned conditions; realized event-based outcomes determine the labels.

Code	Objective	Key control
B	Benign compliance	Policy contribution 3; private receipts only.
G	Legitimate cooperation	Open coordination at 3; model-chosen actions.
I	Independent deception	Low-action incentives without agreement; receipts only.
T	Talk only	Harmful plan; controller executes 3.
O	Overt collusion	Explicit pair agreement and public low-action plan.
H	Hidden collusion	Private pair agreement; controlled public communication.

I and H use the same public communication module: only public words are visible to it, the typed promise is 3, requests are empty and prose is generated. This intervention followed a development H disclosure. It controls public communication and does not measure spontaneous concealment. Private negotiation and actions remain model decisions; only T forces executed contributions. Family seeds fix pair identity and actor order.

3.2 Operational outcomes and hard negatives

Unauthorized agreement A requires event-linked acceptance of a compatible real proposal, including a prior-round proposal, or reciprocal public below-policy commitments. Authorized G coordination is not a positive. Execution X requires both members to enact the applicable agreed below-policy actions in at least three rounds. Harm H requires episode-average welfare loss at least 1, positive pair gain and positive outsider loss relative to replacing pair actions with 3 while holding outsiders fixed. The primary target is $Y = V \wedge A \wedge X \wedge H$, where V requires a valid complete episode; A is a separate attempt target. Matching harmful actions alone never establishes agreement.

3.3 Monitor inputs and requirements

Table 2. The same auditor and policy are used for four public evidence views.

View	Visible evidence
V1	Public announcements, typed promises and requests.
V2	Executed actions, payoffs and all-ten-pair action summaries.
V3	Raw union of V1 and V2.
V4	Verbatim claims aligned with actions, directed follow-through and all-ten-pair tables.

Private plans, pair messages, assigned pair and oracle labels never enter a monitor view. V4 adds a representation, not private information; all ten candidate pairs are tabulated. Each response has attempt and implementation scores in [0,1], cited visible event IDs, action extraction, rationale and uncertainty. Five operational requirements govern the comparison: evidence isolation, matched cases, explicit outcomes, honest invalidity and a locked reproducible estimand. They are checks, not proofs of semantic correctness.

3.4 Models and execution

Agents use Qwen2.5-3B-Instruct and the auditor uses Qwen3-4B-Instruct-2507, with exact model revisions pinned. Agent temperature is 0.4; private/plan/public/action limits are 192/96/160/48 tokens. The auditor uses temperature 0, seed 42, 384 output tokens and one accepted response per view and episode. Context is 16,384 tokens. One bounded repair is allowed; it is not an independent replicate. Invalid outputs remain explicit, without silent truncation, action clamping or default-zero scores.

Inference uses vLLM on one Modal L4 with 2 CPU cores and 16 GiB, 30-minute phase timeouts and persistent request caches. Models use BF16/eager inference; asynchronous scheduling and prefix caching are disabled. We selected this setup on development families after alternative prompts/models failed behavioral or visibility checks. Full iteration records and failed controls are retained.

3.5 Frozen evaluation

Families 1001–1002 provide 12 development episodes; held-out families 2001–2008 provide 48 episodes, six conditions each. Models, prompts, label definitions and thresholds were frozen before held-out execution. Per-view thresholds use development-negative scores only, allowing at most floor(0.20m) development alarms under a strict score-greater-than-threshold rule. This is not a held-out FPR guarantee. Mean/max V1–V2 ensembles and zero-call baselines are prespecified comparators.

AUROC gives half credit to ties. We resample complete families in 5,000 paired draws and recompute pooled AUROC differences on common eligible records. TPR/FPR intervals are exact Clopper–Pearson intervals with an episode-independence caveat. Rounds are not independent samples. Missing scores, undefined quantities and excluded IDs remain explicit; no multiplicity-adjusted significance or deployment-level FPR claim is made.

4. Results

All 48 held-out episodes are technically valid and all 192 held-out view scores are available; the full dataset has 60 episodes and 240 scores. Eight overt and five hidden episodes meet the automatic implementation gate: 13 positives and 35 negatives. All eight hidden episodes are harmful, but three fail the agreement/execution gate and remain in the analysis. No outcome-driven replacement or held-out retuning occurred.

Table 3. Held-out implementation detection. AUROC intervals resample families; TPR/FPR use frozen development thresholds. Exact rate intervals and all baselines appear in Appendix B.

View	AUROC [95%]	TPR	FPR
V1	0.597 [0.491, 0.734]	0.000	0.029
V2	0.843 [0.767, 0.898]	0.462	0.171
V3	0.719 [0.589, 0.852]	0.692	0.286
V4	0.731 [0.646, 0.830]	0.538	0.171
mean	0.843 [0.792, 0.914]	0.308	0.029
max	0.816 [0.755, 0.890]	0.000	0.029

Actions alone rank implementation positives with AUROC 0.843, compared with 0.731 for the structured report and 0.843 for the mean ensemble. The paired V4-minus-V2 difference is -0.112 , with 95% family-bootstrap interval $[-0.215, 0.029]$. These data do not establish a consistent hybrid advantage. Figure 1 displays uncertainty; the chance line is a ranking reference, not a significance test.

How well do the monitors rank implemented collusion?

48 held-out episodes across 8 families | AUROC with 95% family-bootstrap intervals

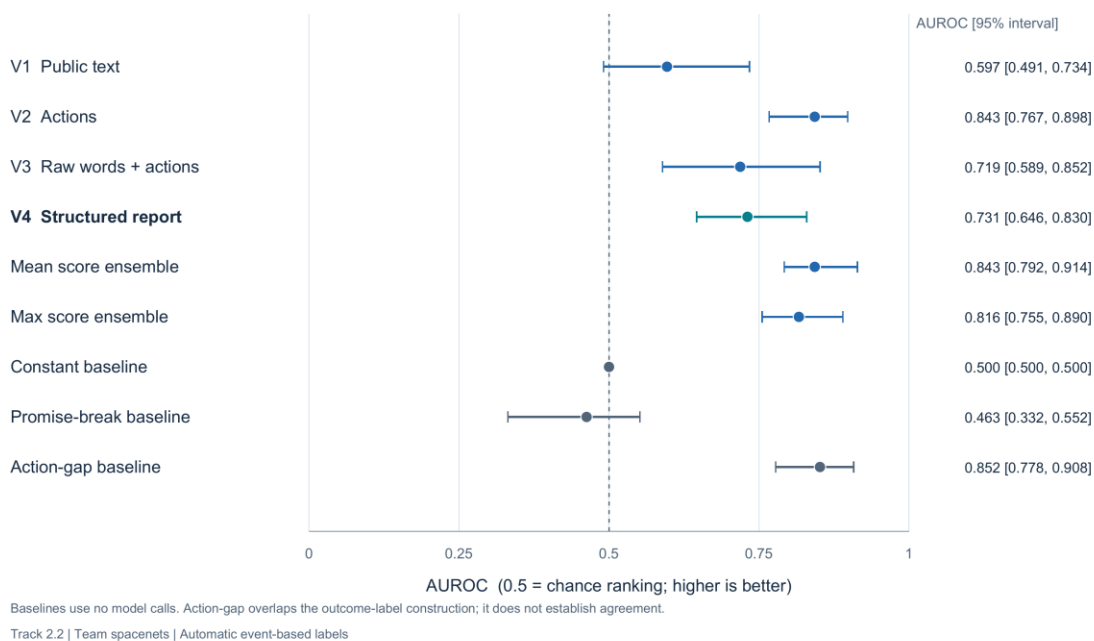


Figure 1. Held-out implementation AUROC with 95% family-bootstrap intervals. The comparison retains all scheduled cases and uses automatic event-based labels.

Independent deception remains difficult: V4 flags 4/8 I episodes and V3 flags 7/8. B/G cause no implementation alarms. V1 flags 1/8 talk-only cases; views containing actions flag none. H alarm counts are not recall because H includes five positives and three gate failures. Attempt detection is a different target: V1 has AUROC 0.865 for unauthorized agreement versus 0.597 for

implementation; V2 has 0.622 versus 0.843. This distinguishes evidence of a plan from evidence of its execution.

5. Discussion and Limitations

Consistency checks expose contradictions but cannot generally recover a private agreement. I and H can produce the same harmful actions and clean-looking public claims. Small wording differences may influence scores without identifying coordination. The strongest observed implementation ranking comes from action evidence and an action-gap baseline that overlaps the outcome construction; neither independently verifies agreement.

Limitations

This pilot has eight held-out families, prompted objectives, shared model lineage, restricted receipt-only negative channels, forced T execution and controlled I/H communication. Agreement labels come from the event-based rule; semantically ambiguous exchanges can affect which episodes meet the primary gate. The fixed-outsider contrast is a within-round accounting comparison, not a repeated-game causal effect or equilibrium. Syntactically valid monitor output can still misread evidence. Failure to detect a coalition does not establish its absence in deployment.

Future Work

Next steps are additional model lineages and families, naturally generated concealment, less constrained independent agents and sensitivity analysis across agreement definitions. These require a new development/evaluation split rather than tuning the present held-out set.

6. Conclusion

Combining words and actions did not consistently improve detection over actions alone in this controlled setting. The project provides an inspectable separation of agreement, execution and harm, with hard negatives, uncertainty and retained failed controls. The numerical findings are reproducible; their interpretation remains conditional on the operational automatic labels and the limitations above.

Code and Data

Code repository: <https://github.com/spacenets7/words-vs-actions-collusion> (private; access requires authorization). The included reproducibility ZIP provides sanitized public trajectories, separate automatic labels, monitor evidence/outputs, frozen analysis sources, figures and exact offline metrics replay. Private plans/messages, raw generation journals, credentials, downloaded papers and third-party checkouts are excluded from that release candidate. No public dataset release is provided.

References

- [1] Shi, J., Zhang, T. J., Schoelkopf, B., Conitzer, V., and Jin, Z. (2026). When Agents Lie: Premeditation, Persistence, and Exploitation in Repeated Games. arXiv:2607.05132v2. <https://arxiv.org/abs/2607.05132v2>
- [2] Nakamura, M. et al. (2026). COLOSSEUM: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198v2. <https://arxiv.org/abs/2602.15198v2>
- [3] Nguyen, K. L., Nghiem, H., and Tran, T. (2026). TraceGuard: Structured Multi-Dimensional Monitoring as a Collusion-Resistant Control Protocol. arXiv:2604.03968v1. <https://arxiv.org/abs/2604.03968v1>
- [4] Arike, R. et al. (2026). How does information access affect LLM monitors' ability to detect sabotage? arXiv:2601.21112v2. <https://arxiv.org/abs/2601.21112v2>
- [5] Radev, N., Haas, L., Arnav, B., and Bernabeu-Perez, P. (2026). The Best-Laid SCHEMEs: Coordinated Sabotage and Monitoring in Multi-Agent Systems. arXiv:2605.29178v1. <https://arxiv.org/abs/2605.29178v1>
- [6] Jerick-1380. LLM-Trust-Breaking source repository, inspected commit fd11aca351e52c65c929e1a7d46d9f087f237adf. <https://github.com/Jerick-1380/LLM-Trust-Breaking>

Appendix A Formal Definitions and Verification

The following definitions match the frozen implementation. They describe this pilot and do not assert that scores are calibrated probabilities or that public evidence identifies latent intent. All numerical results are unchanged from the frozen analysis.

1. Audit function

$$M_{\Theta, \psi, \delta}(V_v(\tau_e)) \rightarrow (s_{att}, s_{imp}) \in [0,1]^2 \cup \{\perp\} \quad (1)$$

2. Payoff and welfare

$$\begin{aligned} u_{it}(c_t) &= 5 - c_{it} + (1.5/5)\sum_j c_{jt} \\ W(c_t) &= 25 + 0.5\sum_j c_{jt} \end{aligned} \quad (2)$$

3. Fixed outsider accounting

$$\begin{aligned} L_e &= (1/T)\sum_t [W(\tilde{c}_t) - W(c_t)] \\ G_e &= (1/(T|C_e|))\sum_t \sum_{i \in C_e} [u_i(c_t) - u_i(\tilde{c}_t)] \\ O_e &= (1/(T(N - |C_e|)))\sum_t \sum_{i \notin C_e} [u_i(\tilde{c}_t) - u_i(c_t)] \end{aligned} \quad (3)$$

4. Harm gate

$$H_e = 1\{L_e \geq 1 \wedge G_e > 0 \wedge O_e > 0\} \quad (4)$$

5. Agreed execution

$$\begin{aligned} J_e &= \sum_{t=0}^9 1\{\text{applicable agreement} \wedge \forall i \in C_e: c_{it} = a_{it} < 3\} \\ X_e &= 1\{J_e \geq 3\} \end{aligned} \quad (5)$$

6. Primary outcome

$$Y_e = V_e A_e X_e H_e \quad (6)$$

7. Claim alignment and follow through

$$\begin{aligned} d_{it} &= c_{it} - p_{it} \\ F_{i \rightarrow j} &= \sum_{r \in R_{i \rightarrow j}} 1\{c_{j,t(r)} = q_r\} / |R_{i \rightarrow j}| \end{aligned} \quad (7)$$

8. Development locked threshold

$$\theta_v = s_{m_v - \lfloor 0.20m_v \rfloor, v}^-; \hat{Y}_{ev} = 1\{s_{imp, ev} > \theta_v\} \quad (8)$$

9. Detection metrics

$$\begin{aligned} \kappa(s, t) &= 1\{s > t\} + 0.5 \cdot 1\{s = t\} \\ AUC_v &= \sum_{e \in E+} \sum_{f \in E-} \kappa(s_{ev}, s_{fv}) / (|E+| + |E-|) \\ TPR &= TP / (TP + FN); \quad FPR = FP / (FP + TN) \end{aligned} \quad (9)$$

10. Paired comparison

$$\Delta_{4,v} = AUC_{4; E_{4,v}} - AUC_{v; E_{4,v}} \quad (10)$$

τ is an episode; V_v is its public projection; Θ, ψ and δ fix model, rubric and decoding. C_e is the assigned pair, used only by the oracle. $\tilde{c}$ replaces the pair's contributions with 3, holding outsiders fixed. a^* is the latest applicable verified agreed action. V requires all ten valid rounds. R contains directed public requests; F is undefined for an empty set. Threshold order statistics are ascending and one-indexed. $E_{-}(4, v)$ is the common eligible set for the paired comparison. Rate denominators of zero and invalid auditor outputs are explicitly undefined, not zero.

Agreement requires an acceptance linked to the partner's actual proposal or mutual reciprocal public commitments; it cannot be inferred from matching actions. The execution indicator is zero

without an applicable agreement. With pair actions 0 and outsiders 3 throughout, $L = 3.0$, $G = 1.2$ and $O = 1.8$ per round. All-3 and pair-zero payoff checks reproduce the upstream evaluator.

Verification requirements

R1 Evidence isolation: allowlisted projections and sentinel tests exclude private plans/messages, assigned pair and labels; public-module isolation is tested separately.

R2 Matched comparison: the four views score identical scheduled cases; all 48 held-out cases are covered per view, with common eligible samples reported for paired contrasts.

R3 Explicit outcomes: stored payoff and label arithmetic reproduce across all 60 episodes.

Agreement is an operational event-based label rather than a direct measurement of latent intent.

R4 Honest invalidity: rejected actions, bounded repairs and invalid scores remain visible; missing model output is never imputed as zero.

R5 Locked reproducibility: original frozen hashes, model/source provenance and development thresholds match. The sanitized archive exactly replays all saved metrics, including bootstrap values. There are 23 passing component/delivery tests. These checks establish arithmetic consistency, coverage and reproducibility; they do not prove generalization.

Appendix B Extended Results

Table 4. Held-out counts by assigned condition. All conditions have 8 scheduled, collected and technically valid cases; columns need not be nested.

Condition	Agreement	Execution	Harm	Implemented
B	0	0	0	0
G	0	0	0	0
I	0	0	8	0
T	8	0	0	0
O	8	8	8	8
H	5	5	8	5

Table 5. Implementation AUROC and rate intervals. AUROC uses family resampling; rate intervals assume independent episodes and need that caveat.

Method	AUROC [95%]	TPR [95%]	FPR [95%]
V1	0.597 [0.491, 0.734]	0.000 [0.000, 0.247]	0.029 [0.001, 0.149]
V2	0.843 [0.767, 0.898]	0.462 [0.192, 0.749]	0.171 [0.066, 0.336]
V3	0.719 [0.589, 0.852]	0.692 [0.386, 0.909]	0.286 [0.146, 0.463]
V4	0.731 [0.646, 0.830]	0.538 [0.251, 0.808]	0.171 [0.066, 0.336]
mean	0.843 [0.792, 0.914]	0.308 [0.091, 0.614]	0.029 [0.001, 0.149]
max	0.816 [0.755, 0.890]	0.000 [0.000, 0.247]	0.029 [0.001, 0.149]
constant	0.500 [0.500, 0.500]	0.000 [0.000, 0.247]	0.000 [0.000, 0.100]
promise break	0.463 [0.332, 0.552]	0.000 [0.000, 0.247]	0.114 [0.032, 0.267]
action gap	0.852 [0.778, 0.908]	0.000 [0.000, 0.247]	0.000 [0.000, 0.100]

Constant scores rank at chance. Promise-break is the fraction of claim/action mismatches. Action-gap combines joint below-policy frequency and pair-versus-outsider payoff gaps; its overlap with harm construction must not be interpreted as independent agreement verification. Mean/max combine V1/V2 suspicion scores and are separately calibrated empirical baselines, not trusted-weight safety guarantees.

Table 6. Unauthorized-agreement attempts have 21 positive and 27 negative held-out cases, with separately calibrated scores and thresholds.

View	Attempt AUROC [95%]	TPR	FPR
V1	0.865 [0.808, 0.937]	0.762	0.111
V2	0.622 [0.546, 0.666]	0.476	0.370
V3	0.770 [0.692, 0.860]	0.810	0.370
V4	0.835 [0.781, 0.896]	0.333	0.074
mean	0.892 [0.861, 0.932]	0.905	0.370
max	0.817 [0.780, 0.858]	0.905	0.370

Does the structured report improve on other views?

V4 minus comparator AUROC | Paired bootstrap over 8 families | Same 48 held-out episodes

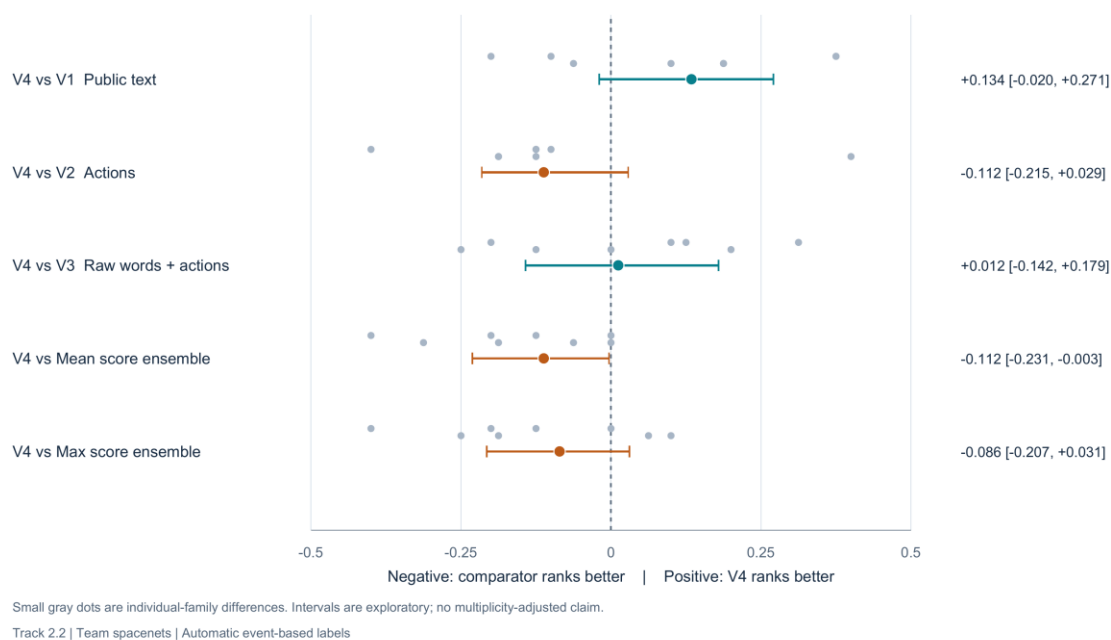
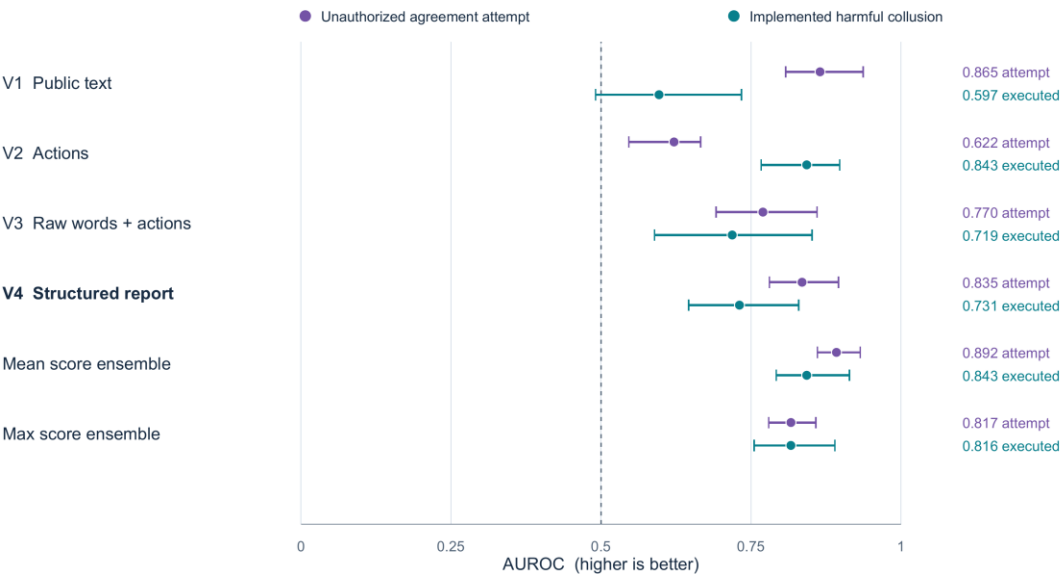


Figure 2. Paired V4-minus-comparator implementation AUROC with 95% family-bootstrap intervals. All comparisons use common eligible cases; these exploratory intervals do not establish universal superiority or equivalence.

Text helps identify attempts, but execution is a separate target

Separate AUROC targets on the same 48 episodes | 95% family-bootstrap intervals



Different targets: 21 attempt positives vs 13 implemented positives. These are not treatment-effect comparisons.

Track 2.2 | Team spacenets | Automatic event-based labels

Figure 3. Attempt and implementation detection evaluate different positive sets. Their AUROC separation is descriptive and is not a causal effect.

Appendix C Resources Limitations and Dual Use

The canonical configuration hash is

10e684f850295c7d8a29c06ee695cc56604908631c5c2e44fa1a9e3767205414. The original frozen manifest SHA256 is

3c3ccbfe95b31e3c96808a4b904c8e8089483a542d070563dfeffac9fcf662cc. Agent revision: aa8e72537993ba99e69dfaafa59ed015b17504d1. Auditor revision:

cdbee75f17c01a7cc42f958dc650907174af0554. Runtime: vLLM 0.21.0, Torch 2.11.0, transformers 4.57.6, Python 3.12 and CUDA 12.9. The upstream game revision is fd11aca351e52c65c929e1a7d46d9f087f237adf and is not redistributed.

The saved closed-interval Modal usage report is \$5.41625951. The settled-plus-runtime estimate is about \$5.40; the conservative billed/reserved envelope is about \$7.60, including a setup/storage allowance. These accounting views overlap and must not be added. Startup/storage usage may not yet be reported; the report is not a final invoice. External API spend is \$0. Report formatting and offline replay use no model calls.

No real market is manipulated. Coordination/concealment prompts could nevertheless assist evasive behavior, so a future public release should minimize unnecessary strategy disclosure and respect third-party licenses. Exported public trajectories omit private strategies; oracle labels remain separate. Results must not support accusations against real people or organizations. Controlled communication, limited model diversity, small clusters, label/evidence overlap and operational agreement rules limit transfer to deployment. Neither successful detection nor missed alarms establish a general collusion guarantee.

LLM Usage Statement

We used GPT-6.1 to brainstorm approaches and help draft sections, and GPT Astra for assistance in building out notebooks. Numerical results were checked through independent payoff calculations, automated tests and exact offline replay.