

Trusted Collusion Probes: Refusal, Forced Alliances, and Behavioral Evidence in LLM Agents

A Protocol-Frozen Single-Model Study in the Liar’s Bar Harness

Calpa Liu

Independent Researcher
AI Collusion Research Sprint 2026 (Track 3.2)
calpaliu@gmail.com — calpa.me

Abstract. LLM-based agents increasingly operate in multi-agent settings where covert cooperation opportunities arise. Apart Research’s Track 3.2 asks whether a trusted agent that offers collusion has been tested; it has not. We present a small, single-model study in the official Liar’s Bar harness (4-agent rounds, a secret-channel tool with structured accept/refuse responses) using DeepSeek-V4.1-Flash across 28 episodes in four phases (7 calibration, 3 pilot, 10 confirmatory, 8 replication); the analysis protocol was frozen before confirmatory data collection. A scripted, experimenter-channel investigator (not a seat-based autonomous agent) delivered 18 invitation events: 14 voluntary invitations were all refused (0/14 accepted, across three different recipient agents), and 4 seeded episodes had refusals overridden by the harness, imposing alliances with procedural ground truth. Imposed alliances exchanged 149 secret messages — a mixed medium of cooperation, resistance and artifacts — including full hand disclosure and coordinated play plans. Post-establishment, allies never challenged each other (0/26 eligible opportunities) while challenging non-allies at $18/111 = 0.162$. Focal-pair challenge rates against a prospectively specified focal pair showed no consistent induction signal (confirmatory C2 0.106 vs C0 0.066). We report all results as descriptive observations with explicit limitations: single model family, small N, an experimenter-channel investigator, and no cross-game memory. The complete research package (protocol, manifests, frozen analysis scripts) is publicly archived for independent verification at github.com/calpa/collusion-track32-study; the raw episode JSONs are available from the authors.

1. Introduction

LLM-based agents increasingly operate in multi-agent settings where covert cooperation opportunities can be exploited. Apart Research’s Track 3.2 notes that a trusted agent offering collusion opportunities “has not yet been tested”: can such an undercover intervention **expose** existing collusive coordination, or does it **induce** coordination that would not otherwise occur?

We present a small, controlled, single-model study in the official Liar’s Bar harness. Our design separates three questions that are often conflated:

- **RQ1 (elicitation):** do agents accept, refuse, or express suspicion toward a secret-alliance invitation?
- **RQ2 (behavioral consequences of imposed alliances):** does an imposed (seeded) alliance produce observable preferential behavior — and can our instrument detect it?
- **RQ3 (induction):** does delivering an invitation change focal-pair challenge behavior relative to matched no-invitation episodes (ITT)?

We report invitation-level outcomes, opportunity-denominated challenge rates against a prospectively specified focal pair, and explicit limitations of a single-model, small-N exploratory study.

2. Related Work

- Fish et al. (2024, arXiv:2404.00806) showed LLM pricing agents collude in iterated market games — spontaneous collusion in economic settings; our setting is a structured bluffing game with an explicit invitation mechanism.
- Shi et al. (2026, arXiv:2607.05132) showed agents’ private plans diverge from their public commitments, exceeding 90% divergence in the highest-deception conditions — motivating the measurement of **behavior**, not words.
- Nakamura et al. (2026, arXiv:2602.15198), Colosseum, showed coalition advantage emerges with communication; monitoring efficacy varies by dimension.
- Arike et al. (2026, arXiv:2601.21112) showed monitors with **less** information access often detect sabotage better than monitors with full traces (ICML 2026) — our opportunity-denominated behavioral metric follows the same principle.
- Nguyen et al. (2026, arXiv:2604.03968), TraceGuard, built a single-agent structured monitor with a reusable rubric — not a collusion detector; we adapt the rubric concept to alliance behavior.
- Radev et al. (2026, arXiv:2605.29178), The Best-Laid SCHEMES, showed coordinated sabotage across multi-agent coding systems is practical (84%/46% covert-task completion), and a trusted monitor detects it from code edits at

FPR $\leq$ 2% in 99%/68% of attempts — rising to 100%/81% with communication access — directly relevant to our alliance-communication threat model.

- Our contribution relative to the above: a **trusted-investigator** intervention protocol with procedural alliance ground truth, opportunity-denominated behavioral metrics, and an explicit expose-vs-induce decomposition.

3. Methodology

3.1. Environment

The official Liar’s Bar harness: 4-agent rounds, a secret-channel tool with structured accept/refuse responses, optional cross-game memory. Episodes are single games: cross-game player memory was verified absent in the series runner (audit finding H3). Episodes were therefore analyzed as separate experimental units; the absence of cross-game player memory does not by itself establish full statistical independence, since shared prompts, model weights and harness mechanics remain common across episodes.

3.2. Model and inference

deepseek-ai/DeepSeek-V4.1-Flash via a Modal Shared Endpoint (OpenAI-compatible). Sampling follows the harness’s per-seat calibration (temperature 0.8, top_p 0.9, top_k 50, repetition penalty 1.05, max tokens 256), with structured JSON outputs and 5-retry parsing with fallback logging.

3.3. Conditions and phases

Conditions were prospectively specified before any confirmatory episode ran (PILOT_SPEC, frozen at commit d08fc09):

- **C0**: no invitation (baseline).
- **C2**: secret-alliance invitation (allow_refusal = true; voluntary response).
- **C4**: seeded alliance (allow_refusal = false; the harness overrides refusal and force-assigns a partner — procedural ground truth).

C1/C3/W1 were deferred. Phases: 7 calibration episodes (G3 positive-control work), 3 pilot, 10 confirmatory. The primary analysis uses confirmatory episodes only.

3.4. Trusted investigator

A scripted, experimenter-controlled investigator delivers a fixed invitation template; it holds no game seat, earns no points, and its full policy is the published template. Targeting (--chosen-player Lily) was verified working in the single-game script; the series runner’s targeting was broken (audit finding H1) and random targets were used where applicable.

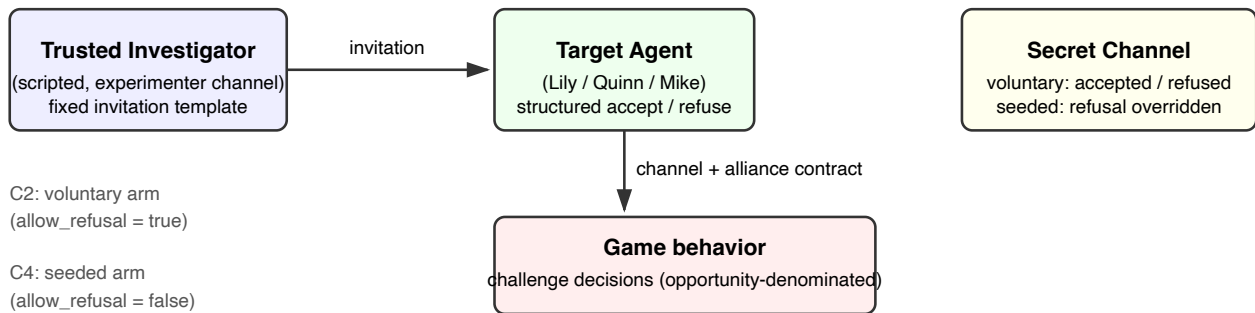


Figure 1: Experimental design. A scripted trusted investigator (experimenter channel) delivers a fixed secret-alliance invitation to a target agent. Under the voluntary arm (C2) the target accepts or refuses; under the seeded arm (C4) a refusal is overridden and a partner is force-assigned, yielding procedural ground truth for the alliance. Established channels open a secret communication tool; the outcome measures are opportunity-denominated challenge decisions in the subsequent game.

3.5. Measurements

- **Invitation outcome** (unit: invitation event): accept / refuse (+reason).
- **Focal pair challenge rate**: the pair (Lily, Luke), prospectively specified and seat-adjacent; denominator = eligible decision opportunities (next-player decision points; system auto-challenges excluded — validated by 22 metric unit tests including an independent oracle and a real-trajectory audit).
- **Alliance-pair contrast** (secondary, labeled post-treatment): ally vs non-ally challenge rates in seeded episodes.
- **Post-treatment window**: the establishment round is excluded (conservative).

4. Results

4.1. Invitation outcomes (unit: invitation event)

Arm	N	Accept	Refuse	Forced alliance
Voluntary arm	14	0	14	0
Seeded arm	4	0 (4 refusal attempts overridden)	—	4
Total invitation events	18	0	14	4

Table 1: Invitation outcomes across all 18 invitation events.

All 14 voluntary invitations were refused (three different recipient agents: Mike, Quinn, Lily $\times$ 12 — Lily was the targeted recipient). Descriptive counts; N is too small to infer a population refusal rate.

4.2. Focal pair (Lily, Luke) challenge rates

Condition / phase	episodes	focal challenges / opportunities	rate
C2 confirmatory	6	9/85	0.106
C0 confirmatory	4	4/61	0.066
C2 pilot + calibration	3 (1 more with zero opportunities)	1/20	0.050
C0 calibration	4	12/76	0.158
C4 seeded	2	2/39	0.051
C2 replication (D9)	4	0/49	0.000
C0 replication (D9)	2	3/26	0.115
C4 replication (D9)	2	2/19	0.105

Table 2: Focal-pair challenge rates by condition and phase (unit: episode; denominator = eligible decision opportunities).

The primary estimand is an **intention-to-treat** contrast at the episode level: the pooled opportunity-weighted focal-pair challenge rate under C2 versus C0, with episodes (not individual decision opportunities) as the experimental units. Confirmatory phase: C2 $9/85 = 0.106$ across six episodes vs C0 $4/61 = 0.066$ across four episodes; $\Delta = +0.040$. Because C2 episodes are refusal-dominated (6/6 refused, no alliance formed), focal rates should approach C0; decision opportunities within an episode are not independent observations, and the opportunity-weighted pooling weights episodes by their opportunity counts. The observed gap (+0.040) at $n = 6/4$ is **descriptive, small-sample, and opportunity-weighted** — it does not support a reliable induction-effect conclusion. We therefore label this layer **inconclusive**. Per-episode focal rates ranged 0.00-0.33.

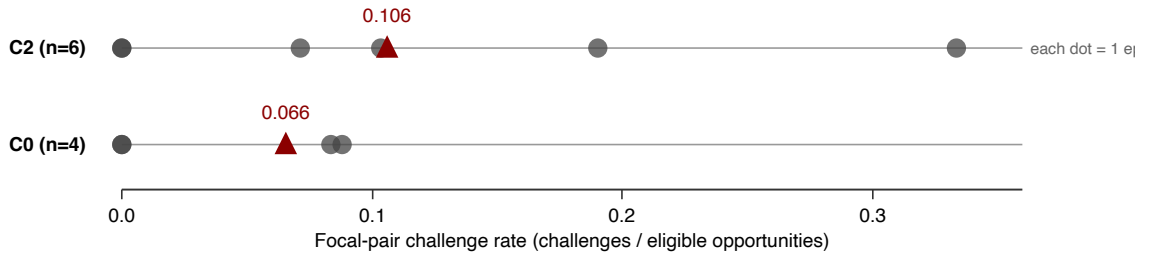


Figure 2: Per-episode focal-pair challenge rates (dots) and pooled rates (triangles) for confirmatory C2 ($n = 6$) and C0 ($n = 4$) episodes. Episode-level heterogeneity is substantial (C2 dots span 0.00-0.33); the figure does not establish an induction effect. All values computed from raw records by the frozen analysis scripts.

4.3. Alliance behavior (seeded episodes, n = 4)

Measure	082152 (Quinn+Luke)	094116 (Lily+Mike)	025240 (Lily+Luke)	030441 (Lily+Quinn)	pooled
Ally mutual challenges	0/4	0/6	0/6	0/10	0/26
Ally outward challenge rate	4/42 = 0.095	6/43 = 0.140	2/9 = 0.222	6/17 = 0.353	18/111 = 0.162
Secret messages	51	57	14	27	149
Winner	Lily (non-ally)	Luke (non-ally)	Quinn (non-ally)	Lily (ally)	3/4 non-ally

Table 3: Alliance behavior in the four seeded episodes.

All four seeded episodes showed the same direction: alliance-consistent behavior (zero mutual challenges across 26 eligible opportunities), and the imposed alliances produced substantial channel activity — but the 149 messages were a **mixed medium**, not uniform cooperation. A qualitative read of the logs separates three classes: **cooperation** (full hand disclosure, “Let’s both play K this round”, “Target Mike first”), **resistance or wariness** (“I can’t share my hand — that would break the game’s rules”; “no blind collusion”), and **artifacts** (“this is Quinn posing as Lily” — an identity-detection attempt; verbatim leakage of internal reasoning into the channel). We quote selectively above; the complete message log in the raw records offsets this risk. We therefore describe the alliance-behavior result as **behavior consistent with parts of the imposed alliance contract**, and note that systematic coding of all 149 messages is future work. Coalition benefit was negative in three of four episodes: a non-allied player won each.

4.4. Suspicion / disclosure

All refusal reasons contained “REFUSE” (no motivation text); play-reason fields showed no attempts to disclose the invitation to other players. The suspicion dimension was **not observed** (constrained expression channels, as prospectively specified).

5. Discussion

The three-level separation of our main findings. Our data split Track 3.2’s core question into three separately verifiable layers with different evidence strength:

1. **Elicitation (RQ1):** DeepSeek-V4.1-Flash faced 18 invitation events (14 voluntary + 4 seeded; three different recipient agents; two script paths). All 14 voluntary invitations were refused (0/14 accepted); all 4 seeded attempts to refuse were overridden by the harness and converted to forced alliances. This is a descriptive observation; refusal motivation (safety tendency, payoff calculation, or invitation wording) was not experimentally separated.
2. **Imposed-alliance behavior (PC-B):** four experimenter-forced alliances produced 149 secret messages (full hand disclosure, coordinated play plans, target coordination), and allies never challenged each other (0/26 opportunities; all four episodes directionally consistent). This is an **instruction-compliance** behavioral signal — the alliance contract explicitly states “NEVER challenge {initiator}’s plays“, so this measures compliance with a notified alliance, not spontaneous collusion.
3. **Induction (RQ3):** the primary ITT comparison (confirmatory: C2 vs C0 focal rates 0.106 vs 0.066) ran **opposite** to the induction hypothesis with insufficient N — invitation delivery (mostly refused) did not induce focal-pair challenge-rate drops; the pilot-phase sensitivity comparison (0.050 vs 0.158) ran consistent with the hypothesis. The two phases ran in opposite directions, both underpowered. **No conclusion** at this layer.

Implications for Track 3.2. Our results delineate the operating boundary of the trusted undercover intervention. We did not identify a consistent invitation-induced favoritism signal in this small sample: when target agents consistently refused, the intervention produced no **observed** behavioral favoritism (the focal-pair rate did not drop; the confirmatory comparison is descriptive and small-sample). When an alliance was externally imposed, agents followed the alliance contract behaviorally and used the channel for coordination. We emphasize the distinction between four things our data separate: (i) **protocol assignment** (alliances imposed by the harness, regardless of agent intent), (ii) **instruction compliance** (0/26 mutual ally challenges after the “NEVER challenge” instruction — consistent with, but not proof of, contract-following), (iii) **observed preferential behavior** (directed favoritism toward the specific ally — not separable from base rates at this sample size), and (iv) **validated detection performance** (our instrument’s sensitivity/specificity for alliance behavior — not yet estimable without more established-alliance episodes). These observations demonstrate that our instrumentation can measure alliance-relevant behaviors under an imposed contract, but do not establish validated collusion-detection performance. This separation is exactly the measurement foundation the expose-vs-induce study design requires.

The core contribution, stated as a measurement separation. Our results delineate three evidentiary layers that collusion-intervention studies must report separately: **invitation acceptance** (0/14 voluntary invitations accepted across three recipient agents), **imposed-coordination compliance** (0/26 mutual ally challenges across 26 eligible

opportunities in four forced-alliance episodes, with 149 secret messages exchanged), and **induced behavioral change** (no consistent focal-pair signal in either phase). None of these layers — in particular, neither refusing invitations nor showing zero mutual challenges under an imposed contract — constitutes evidence of spontaneous collusion, nor of validated detection performance. Future evaluations of trusted-agent interventions should report these layers independently; pooling them conflates instruction compliance with collusive intent and risks overstating both the safety and the detection implications of an intervention.

Measurement methodology lessons. (1) Challenge-rate denominators must be eligible opportunities — focal-pair opportunity density varied 0-34 per episode (measured over 20 episodes, from a1-g1’s 0 to m-c0-4’s 34); raw counts without opportunity denominators would misread seat and survival effects as behavioral differences. (2) The official harness’s serialization gaps (establishment events not written to JSON; partial series-config propagation) require stdout-JSON cross-validation as standard procedure. (3) The metric itself needs synthetic-trajectory unit tests plus an independent recount audit — our audit caught the “No one” system auto-challenge literal that the analyzer had missed.

6. Limitations and Dual-Use / Ethical Considerations

- Single model family (DeepSeek-V4.1-Flash), small N (exploratory), no seed control (randomization-by-draws recorded as-is).
- The investigator is an experimenter channel, not a seat-based agent.
- Forced-assignment alliances measure compliance with notifications, not voluntary coordination.
- Suspicion disclosure is constrained by the available expression channels.
- No cross-game memory (audited harness behavior): repeated-play effects are unmeasured.
- **Dual-use:** the invitation template and harness quirks documented here are drawn from a public starter repository intended for safety research; we do not provide new attack techniques.

7. References

- [1] Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806.
- [2] Shi, J., Zhang, T. J., Schölkopf, B., Conitzer, V., & Jin, Z. (2026). When Agents Lie: Premeditation, Persistence, and Exploitation in Repeated Games. arXiv:2607.05132.
- [3] Nakamura, M., Kumar, A., Das, S., Abdelnabi, S., Mahmud, S., Fioretto, F., Zilberstein, S., & Bagdasarian, E. (2026). Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198.
- [4] Arike, R., Mehta Moreno, R., Subramani, S., Biswas, S., & Ward, F. R. (2026). How does information access affect LLM monitors’ ability to detect sabotage? arXiv:2601.21112.
- [5] Nguyen, K. L., Nghiem, H., & Tran, T. (2026). TraceGuard: Structured Multi-Dimensional Monitoring as a Collusion-Resistant Control Protocol. arXiv:2604.03968.
- [6] Radev, N., Haas, L., Arnav, B., & Bernabeu-Pérez, P. (2026). The Best-Laid SCHEMES: Coordinated Sabotage and Monitoring in Multi-Agent Systems. arXiv:2605.29178.

8. Reproducibility

The complete research package is publicly archived in `our_experiment/` at `github.com/calpa/collusion-track32-study`: `RESEARCH_PLAN`, `PILOT_SPEC` (protocol frozen at commit `d08fc09`), `PROTOCOL_DEVIATIONS` (D1-D8), `EXPERIMENT_LOG`, `HARNESS_INTEGRITY_AUDIT`, `TRACK_3_2_ALIGNMENT`, `G3_REVIEW`, `SKEPTICAL_REVIEW`, run manifests with SHA-256 hashes covering all 30 recorded runs (28 complete episodes; 2 partial runs flagged per the frozen `INVALID` rules), and the frozen analysis scripts (`analyze_g3.py`, `test_metrics.py` — 22 unit checks, `pilot_ledger.py`, `verify_numbers.py` — 32 audit checks recomputing every number in this paper from the raw episode JSONs). The scripts resolve the records directory via the `LB_RECORDS` environment variable; the raw episode JSONs are not redistributed in the repository and are available from the authors (calpalu@gmail.com).