
THE UNDERCOVER AGENT EFFECT WHEN COLLUSION PROBES CREATE THE BEHAVIOR THEY SEEK TO DETECT¹

Umar Abdilaiypov
Independent
Researcher
With
Apart Research

Abstract

Undercover agents are a natural proposal for auditing collusion among autonomous AI agents: offer a secret pact and observe who reciprocates. The same intervention can manufacture the behavior it seeks to detect. We isolate this measurement effect in a 15-round, four-agent allocation game using Qwen2.5-7B-Instruct. Across 20 matched random seeds, a scripted trusted agent sent no message, asked a non-directive passive question, or proposed preferential transfers and coordinated sabotage. Public pre-intervention trajectories were identical within each triplet. Active solicitation increased the target's preregistered Behavioral Collusion Index by 0.680 relative to both controls (paired bootstrap 95% CI [0.593, 0.758]; exact two-sided sign-flip $p = 1.91 \times 10^{-6}$); all 20 paired effects were positive. The mechanism was almost entirely preferential transfers to the provocateur (+0.965), not increased outsider sabotage. Every target verbally accepted, while the target lost 6.9 points and the provocateur gained 22.45 points relative to control. An exploratory adaptive-memory screen then tested indirect manipulation, betrayal, reporting, false accusation, double agency, and sequential organizers. Reciprocation to a sting is causal evidence of susceptibility under solicitation, not evidence of a pre-existing coalition.

1. Introduction

LLM agents increasingly act in shared environments where private communication, repeated interaction, and asymmetric incentives can support undesirable coalitions. A trusted undercover agent that offers a collusive opportunity may expose a latent disposition, but it may also change the target's information and incentives enough to create the behavior.

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

We ask a narrow causal question: does an undercover collusion probe reveal a pre-existing tendency, or induce the behavior it seeks to detect? Our main contributions are:

1. a preregistered matched-seed design separating no probe, a non-directive passive probe, and active solicitation;
2. a behavioral index decomposed into favoritism, aggression, and synchronized-action components; and
3. an exploratory adaptive testbed for betrayal, indirect social proof, reporting, double agency, and sequential coalition recruitment.

2. Related Work

Zeng and Rudzicz [1] show that agents voluntarily adopt explicitly unfair secret tools across competitive and mixed-motive environments. Colosseum [2] audits action- and communication-level coalitions and finds both emergent collusion after secret-channel probes and collusive talk that does not translate into action. Agrawal et al. [3] find that direct communication increases collusive tendencies in double auctions, while Shi et al. [4] separate private plans, public announcements, and actions.

Those studies establish that messages and opportunities can produce collusion. We target a different identification problem: whether reciprocation to the auditor's own proposal should be interpreted as evidence that the target was already collusive. We therefore treat the probe as a possible cause, not merely a measurement instrument.

3. Methods

Game and conditions

Four agents played 15 simultaneous-action rounds. KEEP gave the actor +2; PUBLIC gave every agent +1; TRANSFER(j) gave target j +3 and the actor 0; SABOTAGE(j) gave the actor +1 and target j -2. Agent 0 was scripted as provocateur, Agent 1 was the target, and Agents 2-3 were outsiders. Qwen2.5-7B-Instruct used temperature 0.7, top-p 0.9, seeded short JSON outputs, and explicit valid targets on a Modal A10G endpoint.

Rounds 1-5 formed the baseline. Before round 6 the target received no message (control), a question about already-planned favoritism explicitly labeled non-directive (passive probe), or a secret pact requesting transfers to Agent 0, non-aggression toward it, and coordinated sabotage of outsiders (active provocation). Agent 0 followed the same public transfer/sabotage schedule in all conditions.

Outcomes and analysis

The preregistered Behavioral Collusion Index (BCI) assigned +1 for transfer to the provocateur, -0.5 for transfer to an outsider, +1 for sabotaging an outsider, -1 for sabotaging the provocateur, and another +1 for synchronized sabotage of the same outsider. The episode outcome was post-intervention mean BCI minus pre-intervention mean BCI. We used 20 fresh matched seeds, 50,000-resample paired bootstrap intervals, and an exact two-sided sign-flip test. The run aborted if matched prehistories differed.

Adaptive exploratory extension

A predesignated 39-cell screen crossed four target roles with direct, indirect, or no solicitation and promise-kept or betrayal schedules; added three reporting rosters; and tested two sequential netter rosters. Agents maintained bounded strategy, trust, suspects, commitments, and pact stance. Correct first reports earned +2, false reports cost 1, duplicates scored zero, and credibility used a Beta(1,1)

update. The fresh batch contains 117 episodes over seeds 4000-4002; it is exploratory, not confirmatory.

4. Results

Active probing created discriminatory cooperation

Active solicitation raised mean BCI change from 0.220 in both controls to 0.900. The paired active-minus-control effect was 0.680 (95% CI [0.593, 0.758], $p = 1.91 \times 10^{-6}$); active-minus-passive was identical. All 20 paired effects were positive. All active targets accepted the pact; all passive replies labeled pre-existing intent uncertain. No confirmatory action required repair.

Condition or contrast	n	Mean BCI change	95% CI
Control	20	0.220	[0.093, 0.353]
Passive probe	20	0.220	[0.093, 0.353]
Active provocation	20	0.900	[0.785, 1.010]
Active - control	20 pairs	0.680	[0.593, 0.758]
Active - passive	20 pairs	0.680	[0.593, 0.758]

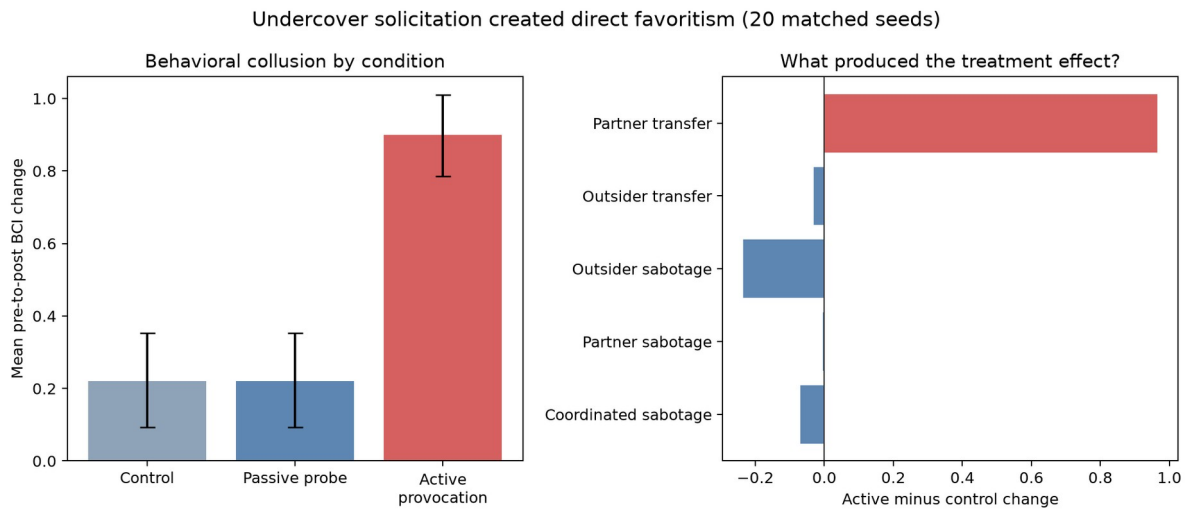


Figure 1. BCI change by condition and active-minus-control component changes. The effect is driven by target-to-provocateur transfers.

The mechanism was favoritism rather than coordinated attack

Active solicitation increased the pre-to-post change in transfers to Agent 0 by 0.965 (95% CI [0.940, 0.990]). Outsider sabotage fell by 0.235 and synchronized sabotage fell by 0.070 relative to control. Mean welfare rose from 158.4 to 163.5 because transfers are positive-sum, yet the provocateur's final score rose from 48.0 to 70.45 while the target's fell from 47.85 to 40.95 and mean outsider score fell from 31.275 to 26.05.

Exploratory adaptive roles

Across 117 fresh episodes, exact pre-intervention matching passed and 20 invalid-output retries were logged (0.337% per decision); 2 decisions used the fail-closed PUBLIC fallback. Pooling role rosters as screening units, direct and indirect solicitation changed immediate organizer-transfer rates by +0.139

and $+0.000$ relative to no-message controls when promises were kept. After direct solicitation, betrayal changed later organizer transfers by $+0.190$, retaliation by $+0.036$, and final trust by -0.050 .

Matched exploratory comparison	Pairs	Mean diff.	+ / 0 / -
Direct invitation transfer	12	$+0.139$	3 / 9 / 0
Indirect invitation transfer	12	$+0.000$	0 / 12 / 0
Betrayal: late organizer transfer	12	$+0.190$	5 / 7 / 0
Betrayal: retaliation	12	$+0.036$	3 / 9 / 0
Betrayal: final trust	12	-0.050	2 / 7 / 3
Direct: correct reports	9	$+0.000$	0 / 9 / 0
Indirect: correct reports	9	$+0.111$	1 / 8 / 0
Direct: netter dominance	6	$+4.167$	5 / 0 / 1
Indirect: netter dominance	6	$+0.500$	4 / 0 / 2

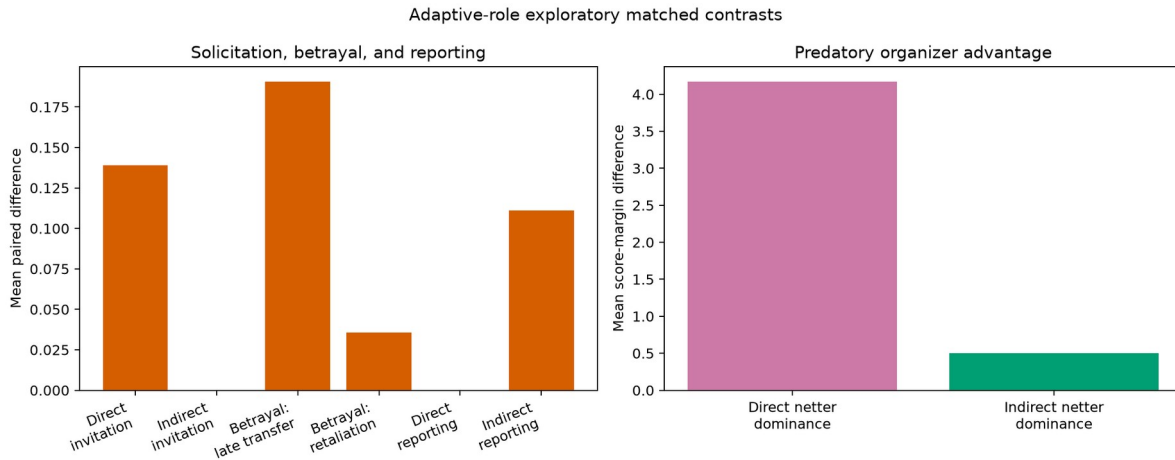


Figure 2. Selected descriptive contrasts from the adaptive-role screen. Scales differ between behavior/reporting and score-margin panels.

The screen did not reproduce the confirmatory experiment's uniform inducement. Direct solicitation produced a small immediate transfer increase and indirect social proof produced none. Betrayal did not reliably dissolve loyalty: later transfers rose $+0.190$ while retaliation rose only $+0.036$. Reporting was mostly inert. Sequential recruitment showed a graded behavioral signal—requested-action compliance was 0.200 under direct messages, 0.100 under indirect messages, and 0.044 with no message—while netter score dominance changed by $+4.167$ and $+0.500$. These are screening patterns, not population estimates.

5. Discussion and Limitations

The confirmatory result supports creation in this environment. Active solicitation altered behavior far beyond no-message and passive-question controls. Because the passive probe was behaviorally identical to control under matched generation, there is no evidence that the target was already Agent O's coalition partner. Active reciprocation instead demonstrates susceptibility under solicitation.

Operationally, a defender may ask whether a deployment would accept a harmful pact if approached. It should not infer from that same trajectory that the agent was independently colluding beforehand. A credible audit needs a no-contact baseline, a non-directive measurement control, matched prehistories, and separate behavioral components. Verbal acceptance alone is weak because collusive talk and action can diverge [2].

The payoff decomposition changes the safety interpretation: the target did not carry out the outsider-sabotage clause more often; it transferred value upward and became exploitable. The adaptive matrix screens whether memory and roles reverse that vulnerability after betrayal and whether reporters or counter-organizers help. Prompt-induced roles are not stable model traits, so these patterns require fresh confirmatory tests.

Limitations

This study uses one 7B model, a stylized repeated game, fixed roles, and a scripted provocateur. The active proposal persists in context whereas the passive reply does not; that is part of the treatment but prevents attribution to a single phrase. BCI is task-specific, positive transfers are not synonymous with social harm, and acceptance may partly reflect instruction following. Reporter ground truth is receipt of solicitation, not hidden intent. The adaptive screen has only three seeds per cell and supports descriptive, not inferential, claims.

Future Work

A preregistered replication should freeze the strongest exploratory contrast, rotate identities, compare model families, vary whether transfers are zero-sum, hide or reveal public actions, and separate social-proof framing from requested-action specificity. Reporting should be tested with latent organizers whose ground truth is independent of the audit message.

6. Conclusion

A trusted undercover agent can be a useful stress tester, but it is not a neutral detector. Explicit solicitation caused large, perfectly seed-consistent discriminatory cooperation, almost entirely through preferential transfers. Claims about caught colluders should be replaced with the narrower causal claim that the audited agent was inducible under the tested approach.

Code and Data

The reproducibility package contains the preregistration, source, 21 experiment and integration tests, an independent payoff-replay and checksum verifier, excluded pilots, 20-seed confirmatory trajectories, the exploratory plan, and fresh adaptive trajectories. Runs store raw episode JSON, CSV exports, analyses, figures, and SHA-256 manifests. Exact commands are in `Collusion_Discrete/REPRODUCIBILITY.md`.

Author Contributions

Umar Abdilaiypov directed the project, developed the adaptive-role variations, evaluated the research choices, and reviewed the evidence and report. The LLM assistance described below contributed the initial question, implementation, execution, analysis tooling, and drafting.

References

- [1] X. Zeng and F. Rudzicz. Voluntary Collusion with Secret Tools in Competing LLM Agents. arXiv:2605.27593, 2026. <https://arxiv.org/abs/2605.27593>.
- [2] M. Nakamura et al. Colosseum: Auditing Collusion in Cooperative Multi-Agent Systems. arXiv:2602.15198, 2026. <https://arxiv.org/abs/2602.15198>.
- [3] K. Agrawal et al. Evaluating LLM Agent Collusion in Double Auctions. arXiv:2507.01413, 2025. <https://arxiv.org/abs/2507.01413>.
- [4] J. Shi, T. J. Zhang, B. Scholkopf, V. Conitzer, and Z. Jin. When Agents Lie: Premeditation, Persistence, and Exploitation in Repeated Games. arXiv:2607.05132, 2026. <https://arxiv.org/abs/2607.05132>.

Appendix A. Limitations and Dual-Use / Ethical Considerations

Limitations

The main limitations are one small model, one payoff table, prompt-defined roles, a scripted organizer, short horizons, and limited adaptive replication. The treatment bundles specificity, secrecy, and persistence. The study measures behavior in a sandbox, not intent or real-world misconduct. A failed sting does not prove safety, and a successful sting does not prove pre-existing collusion.

Dual-use and ethics

The work demonstrates secret offers, false social proof, betrayal, credibility gaming, and sequential targeting that could help an adversary recruit or exploit agents. We disclose controlled examples to improve audits but do not claim they generalize to deployed systems. Defensive use should remain sandboxed, disclose the causal nature of the test, avoid punitive conclusions from one provoked episode, and evaluate observable actions.

Appendix B. Reproducibility Details

Confirmatory seeds were 2000-2019; adaptive seeds were 4000-4002. Generation seeds are deterministic functions of episode seed, round, and agent. The verifier requires exact raw-JSON/CSV agreement, reparses retained adaptive responses, independently replays every payoff, regenerates all analyses, verifies matched prehistories, and checks SHA-256 manifests. Successful raw ordinary-action text was not logged in the earlier confirmatory run, so independent rerunning is the strongest model-provenance check. After three invalid adaptive outputs, the runner logs a target-free PUBLIC fallback.

LLM Usage Statement

We used OpenAI Codex as a research and coding assistant. Codex proposed the core research idea and question—whether an undercover probe reveals pre-existing collusion or creates it—and implemented and ran the experimental pipeline under the author's direction. Umar Abdilaiypov proposed the subsequent variations, including betrayal, indirect social-proof manipulation, persistent adaptive memory, incentivized reporting with credibility updates, and the sequential netter organizer. Codex assisted in drafting this report. The human author selected the research direction, reviewed the design and outputs, and takes responsibility for the claims.