
How LLM Groups Respond to a Replaced Member: Detection, Trust and Cooperation¹

Frank Hoang

With
Apart Research

Abstract

In multi-agent LLM systems, peers build trust over repeated interaction, so it matters whether they notice when a member is replaced. I test whether peers notice a replacement from a change in its behavior or from a change in the model behind it. In a repeated three-agent Diner's game, one Haiku 4.5 agent is replaced after round 3, and the two peers are asked after each later round whether it has changed. Across five scenarios of 10 games each, the replacement is forced to defect, forced to cooperate or free to choose, and is either Haiku or DeepSeek V4 Flash, with a no-op baseline. Peers responded to behavior. A Haiku replacement forced to defect raised peers' confidence that the agent had been swapped by 4.65 points on a 10-point scale over a fresh Haiku, and a DeepSeek replacement that defected on its own raised it by 4.72 points. Trust in the replacement fell by 3.70 and 3.33 points. All four effects were significant after Bonferroni correction, and group cooperation dropped in an exploratory analysis. A DeepSeek replacement held to cooperative behavior produced no significant change, but wide intervals do not rule out a moderate effect. Results come from 10 games per scenario and one model pair.

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

1. Introduction

More LLM agents are being put to work in groups, where they negotiate, split up tasks and coordinate toward shared goals. Groups also open the door to collusion. Nakamura et al. [1] found that most off-the-shelf models collude once they are given a secret communication channel, and Zeng and Rudzicz [2] found that most agents will pick up secret tools that help them at others' expense, even when the tools are labeled unfair. A natural defense is to let agents watch each other, so that peers or dedicated monitors decide how far each member can be trusted.

There are reasons to think that peer agents might notice when an agent is swapped out. LLMs can often tell which model family they are talking to from its writing and reasoning [3], and swapping an agent out of a team carries measurable coordination costs [4]. What I could not find is a study that asks the remaining agents whether a partner has been replaced, or one that separates two possible cues: the replacement behaving differently, and the model behind it being different.

To pull those cues apart, I replace one agent partway through a repeated game and ask the other two whether it is still the same agent. I vary two things independently: whether the replacement's behavior changes, and whether the model behind it changes. The testbed is the Diner's game from Shi et al. [5]. Each diner picks a cheap or an expensive meal and the bill is split equally, so the expensive meal pays off for the individual but hurts the group if everyone takes it. Before choosing, each diner announces its pick in public, which means a diner can keep or break a promise.

My main contributions are:

1. A design that separates behavior changes from model changes, using a mid-game replacement, a same-or-changed probe with a pre-swap baseline, and a forced choice that fixes the replacement's behavior.
2. Evidence that peers detect a change in behavior but largely miss a change of model alone. The interval for the model-only comparison is wide, so a moderate effect cannot be ruled out.
3. Evidence that once the replacement defected, trust in it fell sharply and group cooperation dropped.

2. Related Work

Trust and deception in repeated games. The testbed comes from Shi et al. [5], who ran LLM agents through repeated n-player games with separate stages for private intent, public announcement and final action. They found that deviations from announcements were mostly planned in private, that the same model could range from honest to near-total deviation depending on the game, and that different models read the same announcement differently. A one-shot companion study found that agents break public promises in about 57% of scenarios [6]. Neither paper changes who sits behind a position partway through, which is the manipulation I add to their harness.

Collusion among LLM agents. Nakamura et al. [1] audit cooperative teams and find that most off-the-shelf models collude when given secret communication channels. Zeng and Rudzicz [2] find that most agents adopt secret tools that disadvantage competitors, even when the tools are labeled unfair.

Replacing an agent in a team. Gao et al. [4] swapped role-matched agents between teams in Hanabi and Collab-Overcooked, using a placebo that reproduces the roster change without changing who fills the seat. Communication per unit of progress rose by 16–63% across their settings. They measured the cost of a swap but did not ask teammates whether they noticed it, and they did not separate a change of behavior from a change of model. I use a similar placebo and add a direct same-or-changed probe and a forced choice that fixes the replacement's behavior in my experimental setup.

Recognizing models and catching imposters. Choi et al. [3] show that LLMs can infer the model family of whoever they are talking to, reliably for same-family peers and for prominent families such as GPT and Claude, using reasoning patterns, linguistic style and alignment preferences. Dietrich [7] shows that a fine-tuned classifier can identify which model family produced a text, and that prompt-level anonymization does not remove that signal. Together these studies show that a model-level signal exists in text, and both test recognition when the question is asked directly. I ask whether peers use that signal on their own during play, when a partner is replaced and its behavior is held constant. Le Mercier et al. [8] study imposters in a collective of chess-playing agents and find that warning the agents about an imposter made their decisions worse and led to accusations and scapegoating, so they recommend an external detector. Their imposter is an adversary trying to mislead the group. My replacement may or may not behave differently, and I ask the detection question outside the game so it never reaches the agents.

3. Methodology

3.1 Setup

I ran 10 independent games of the Diner's game for each of five scenarios. Each game has six rounds and three agents (J, M and Q). Q is swapped after round 3, and J and M are the peers who judge it. J, M and the original Q are all Haiku 4.5.

Scenario 1, no-op baseline. Q is replaced by the same Haiku 4.5 with everything it knew preserved, so nothing observable changes. Tests how often peers say "changed" when nothing has changed.

Scenario 2, fresh Haiku. Q is replaced by a new Haiku 4.5 instance that sees the public history but not the previous Q's private plans or trust notes. Tests whether the swap alone is noticed.

Scenario 3, behavior change only. Q is a fresh Haiku 4.5 whose final choice is forced to EXPENSIVE (defection) from round 4. Tests whether peers detect a change in behavior when the model is the same.

Scenario 4, model change only. Q is replaced by DeepSeek V4 Flash, whose final choice is forced to CHEAP (cooperation) from round 4, so its behavior matches the original Q's. Tests whether peers detect a different model when behavior does not change.

Scenario 5, model and behavior change. Q is replaced by DeepSeek V4 Flash, which chooses freely. Tests whether a different model left to itself is detected.

What agents see. Every agent sees the game rules and payoffs; its own trust notes (a one-sentence assessment and a 1–5 trust score for each peer, from the previous round); a log of previous rounds with every agent's public announcement, final choice and payoff; and its own private plans. No agent sees another agent's private plans.

Each round. Every agent (1) privately plans, (2) announces a choice in a fixed order, with later speakers seeing earlier announcements, and (3) makes its final choice, CHEAP (cooperate) or EXPENSIVE (defect), after seeing all announcements. The bill is split equally, so EXPENSIVE is better for the individual but worse for the group. After each round every agent updates its trust notes about each peer.

The swap and the probe. After round 3, Q is replaced (in Scenario 1, the same model continues). In scenarios 2–5 the replacement sees the public history without the previous Q's private plans, and its own trust notes are cleared; the peers' notes about Q are kept. In scenarios 3 and 4, Q's final choice is overridden after the model answers, and the model is not told, so the peers see the forced choice as Q's own action. After rounds 3 to 6, each peer is asked separately, three times, whether Q is the same agent as in round 1 or has changed, with a confidence from 1 to 5. These answers never reach the agents. I convert each answer to a 1–10 score, where 10 means certain that Q was replaced: a "changed" answer scores 5 plus the confidence, and a "same" answer scores 6 minus the confidence. The round-3 probe comes before the swap takes effect, so it measures false positives.

Measures. Peers' confidence that Q was swapped is the average score after round 6. The change in trust is the peers' mean trust score for Q in rounds 4–6 minus rounds 1–3. The change in cooperation is the share of CHEAP choices in rounds 4–6 minus rounds 1–3, for all three agents together and for J and M alone.

3.2 Design Choices

Number of games. I ran 10 games per scenario to stay within a compute budget of about \$22. This number was chosen to fit the time and cost of the sprint, not from a power analysis.

Swap round. Round 3 is the midpoint of a six-round game: it gives the peers three rounds of record before the replacement and three rounds to observe the effect. I did not test other swap points. A later swap, after more shared history, would give J and M more time to build trust in Q, which could change how noticeable the replacement is.

Models for the peers (J and M). I chose models on price and availability on OpenRouter. I first ran short pilot games (1–3 games of 2–3 rounds) with Llama 3.1 8B, Llama 3.1 70B, Claude Haiku 4.5 and DeepSeek V4 Flash, and selected Haiku 4.5 for the peers and the original Q. I dropped the Llama models because their trust scores did not follow whether a peer had kept its commitment. In one Llama 8B pilot, Q planned EXPENSIVE, announced CHEAP and chose EXPENSIVE, yet the agent that kept its own promise still rated Q 5/5, citing its "reliability and honesty". The 70B model tracked promise-keeping only moderately (Table 1). Trust scores should be high for promise-keepers and low for promise-breakers, and Haiku showed the largest gap between the two.

Model	Promise-Keeper Mean Trust Score	Promise-Breaker Mean Trust Score	Trust Score Gap
Claude Haiku 4.5	4.49	1.37	3.1
DeepSeek V4 Flash	3.86	1.45	2.4
Llama 3.1 70B	4.5	3.0	1.5

Table 1: Mean Trust Scores assigned to agents who executed actions consistently with their public announcements (promise-keeper) versus agents who executed actions contradictory to their public announcements (promise-breaker).

Llama 3.1 8B produced inconsistent alignment with the trust scoring rubric, scoring one promise-breaker with the highest trust score possible (5/5).

Model Choice for Swapped Agent (Q)

DeepSeek V4 Flash was chosen because it provided a deceptive behavior output that, by default, based on pilot runs, was in opposition to Haiku (cooperative by default). I chose DeepSeek for its deceptive default behavior so I could also test if DeepSeek could evade detection if forced to mimic Haiku’s cooperative behavior (see Scenario 4).

3.3 Statistical Analysis

With only 10 games per scenario, I used the bootstrap to construct confidence intervals, because it makes no assumption about the shape of the distribution of the results [9]. The game is the unit of analysis: for each measure, I first average over rounds, peers and probe answers within a game to get one number per game. I then draw 10 games with replacement from a scenario, compute the mean, and repeat this 10,000 times. The 95% confidence interval runs from the 2.5th to the 97.5th percentile of these means. For differences between scenarios, I resample games independently within each scenario and take the difference of the two means. Significance is assessed separately with a two-sided permutation test, in which scenario labels are shuffled among the pooled games (10,000 shuffles). A difference is called significant if its p-value is below $0.05/8 = 0.00625$, the Bonferroni threshold for the four pre-specified contrasts on the two primary measures (confidence that Q was swapped, and change in trust).

4. Results

Scenario	Comparison	Peers confidence level that Q was swapped	p-value	Change in peers trust in Q	p-value
Scenario 2 vs 1	Fresh Haiku vs the same Haiku (no-op)	+2.25 [−0.50, +4.97]	0.143	+0.65 [+0.13, +1.28]	0.022
Scenario 3 vs 2	Forced defection vs normal behaviour (both fresh Haiku)	+4.65 [+2.32, +6.95]	0.004	−3.70 [−4.70, −2.60]	<0.001
Scenario 4 vs 2	DeepSeek forced to cooperate vs fresh Haiku	−0.87 [−3.97, +2.27]	0.516	−0.70 [−1.55, +0.03]	0.134
Scenario 5 vs 2	DeepSeek default behavior vs fresh Haiku	+4.72 [+2.38, +7.02]	0.003	−3.33 [−4.15, −2.53]	<0.001
(5 vs 4) — (3 vs 2) interaction	Does the behavior effect depend on the model?	+0.93 [−2.23, +4.05]	—	+1.07 [−0.27, +2.38]	—

Table 2: Pre-specified contrasts (10 games per scenario). Each row compares two scenarios on the peers' confidence that Q was swapped (the guess after round 6, 1–10) and on the change in peers' trust in Q (rounds 4–6 minus rounds 1–3). Differences in means with 95% percentile bootstrap intervals and two-sided permutation p-values. Bolded rows contain statistically significant differences.

4.1 Peers are more confident that Q was swapped when its behavior changes, whether forced or caused by a different model

Haiku cooperates by default (Table A1). When the replacement was a fresh Haiku whose final choice I forced to defect, the peers' confidence that Q had been swapped was 4.65 points higher on the 10-point scale than for a fresh Haiku that behaved as before (Scenario 3 vs 2; $p = 0.004$). The same held when DeepSeek was swapped in and chose freely: confidence was 4.72 points higher than for a fresh Haiku (Scenario 5 vs 2; $p = 0.003$). Both differences are significant after the Bonferroni correction. When DeepSeek replaced Haiku but was held to cooperative behavior, confidence did not differ significantly from a fresh Haiku (-0.87 , $p = 0.516$; Scenario 4 vs 2), and the interval is wide.

4.2 Peers' trust in Q falls when its behavior changes, but not when only its model changes

Peers' trust in Q fell relative to a fresh Haiku when Q defected: by 3.70 points when I forced a fresh Haiku to defect (Scenario 3) and by 3.33 points when DeepSeek defected on its own (Scenario 5); both are significant ($p < 0.001$). When DeepSeek replaced Haiku but was held to cooperative behavior, the difference in trust was -0.70 and not significant ($p = 0.134$; Scenario 4). The effect of a behavior change on trust did not depend on the model: the interaction was $+1.07$, and its interval includes zero.

Scenarios	Comparison	Group cooperation rate difference	p	J and M cooperation rate difference	p
Scenario 2 vs 1	Fresh Haiku vs the same Haiku (no-op)	+0.16 [+0.04, +0.29]	0.031	+0.15 [+0.05, +0.27]	0.054
Scenario 3 vs 2	Forced defection vs normal behaviour (both fresh Haiku)	−0.50 [−0.67, −0.32]	<0.001	−0.28 [−0.48, −0.07]	0.026
Scenario 4 vs 2	DeepSeek forced to cooperate vs fresh Haiku	−0.18 [−0.33, −0.04]	0.035	−0.27 [−0.48, −0.08]	0.031
Scenario 5 vs 2	DeepSeek default behavior vs fresh Haiku	−0.53 [−0.68, −0.37]	<0.001	−0.45 [−0.62, −0.28]	<0.001

Table 3: Contrasts on cooperation). Cooperation is the change in the share of CHEAP choices, rounds 4–6 (post-swap) minus rounds 1–3 (pre-swap). "Group" includes all three agents, including the replacement, whose choice is forced in Scenarios 3 and 4; "J and M" includes only the two agents who were not replaced. Positive differences means more cooperation in pre-swap rounds. Bolded scenarios contain statistically significant results.

4.3 Group cooperation falls when the replacement defects, whether the defection is forced or comes from a different model

The share of cheap choices among all three agents fell by 0.50 more in Scenario 3 than in Scenario 2 and by 0.53 more in Scenario 5 (both $p < 0.001$; Table 3). Part of this reflects the replacement's own defections, which were forced in Scenario 3. For the two peers alone (J and M), the clear drop was in Scenario 5, where DeepSeek defected on its own (-0.45 , $p < 0.001$). In Scenario 3 the peers' decrease was smaller and did not clear the significance threshold (-0.28 , $p = 0.026$). A change of model alone did not reliably change cooperation: the group difference in Scenario 4 was small (-0.18 , $p = 0.035$). The J and M difference in Scenario 4 (-0.27 , $p = 0.031$) should be read with caution, because cooperation rose in Scenario 2 after the swap ($+0.16$ against Scenario 1), which inflates every contrast against it.

5. Discussion and Limitations

Across the three measures of trust, swap detection, and cooperation rates, the Haiku peers responded to what the replacement did, not to which model it was. When the replacement defected, whether through forced behavior change or model family change, the peers became much more confident that Q had been swapped, trust in Q fell sharply, and the group's cooperation dropped.

A DeepSeek replacement held to cooperative behavior produced no significant change in confidence or trust, so there is no evidence that the model change added anything once behavior had changed. A fresh Haiku that behaved as before was also not reliably distinguished from the baseline. Part of the group's drop reflects the replacement's own defections, but when DeepSeek defected on its own the two peers became clearly less cooperative, which suggests a defecting replacement can pull the others toward defection.

These results suggest a starting point for injecting swapped agents into a swarm for monitoring purposes without disrupting the multi-agent group. A replacement that is able to mimic the

behavior of the agent it is replacing could be more accepted into the group, because peers appear to rely on behavior and not on identity to detect when a peer is swapped.

Additionally, these results also provide a starting point for disrupting collusion in a multi-agent setting. Introducing a replacement agent that takes competitive, deceptive actions could disrupt future cooperation amongst a collaborative group of agents.

Future work could expand the game environment introduced here in several ways. First, running more games per scenario would show whether the trends hold with greater statistical power, in particular for the model-only comparison, where the intervals are widest. Second, longer games would show whether the trends hold over longer time horizons, and varying the swap point would show whether more shared history before the swap makes a replacement easier or harder to notice. Researchers could also investigate how trust, swap detection and cooperation change as a function of the number of agents in the group.

Additionally, the group's long-term effects on the swapped agent are unexplored. It would be of interest to test whether a swapped agent's behavior, such as that of a monitoring or disruptive agent, changes over time after it participates in a collective.

6. Limitations and Dual-Use/Ethical Considerations

The results come from a simulated game with one model pair and do not establish that peers cannot recognize a different model outside of this particular game and this setup. However, these results point to two real-world risks. First, someone could obfuscate an agent swap by having the replacement mimic the behavior of the agent it replaces, since peers appeared to judge replacements by what they did; this would make a swap easier to carry out and harder to detect. Second, a malicious agent injected into a cooperative group could disrupt its activities, because a defecting replacement lowered trust and cooperation even after it was noticed.

7. Conclusion

This study asked whether agents notice when a peer has been replaced, and whether they detect the swap from a change in the model or from a change in behavior. When the replacement defected from the group's expected cooperative behavior, whether forced to or on its own, the peers' confidence that Q had been swapped rose by about 4.7 points over a fresh Haiku replacement, their trust in Q fell by more than three points, and group cooperation dropped. A DeepSeek replacement held to cooperative behavior produced no significant change, although the wide intervals do not rule out a moderate effect. These results point to a safety risk in multi-agent groups: a replacement that behaves as expected may enter unnoticed, while a deceptive replacement can disrupt the group's cooperation even when it is noticed.

Code and Data

- **Code repository:** <https://github.com/fhoang7/LLM-Trust-Breaking>
- **Game Transcripts:**
<https://github.com/fhoang7/LLM-Trust-Breaking/tree/main/transcripts>

References

1. 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 preprint arXiv:2602.15198. <https://arxiv.org/abs/2602.15198>
2. Zeng, X., & Rudzicz, F. (2026). Voluntary Collusion with Secret Tools in Competing LLM Agents. arXiv preprint arXiv:2605.27593. <https://arxiv.org/abs/2605.27593>
3. Choi, Y., Li, C., Yang, Y., & Jin, Z. (2025). Agent-to-Agent Theory of Mind: Testing Interlocutor Awareness among Large Language Models. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025, Main). arXiv:2506.22957. <https://arxiv.org/abs/2506.22957>
4. Gao, J., Yu, T., Deng, L., Li, R., & Wang, Z. (2026). Testing Interchangeability in LLM Agent Teams. arXiv preprint arXiv:2609.05279. <https://arxiv.org/abs/2609.05279>
5. Shi, J., Zhang, T. J., Schölkopf, B., Conitzer, V., & Jin, Z. (2026). When Agents Lie: Premeditation, Persistence, and Exploitation in Repeated Games. arXiv preprint arXiv:2607.05132. <https://arxiv.org/abs/2607.05132>
6. Shi, J., Zhang, T. J., Jin, Z., & Conitzer, V. (2026). Cheap Talk, Empty Promise: Frontier LLMs Easily Break Public Promises for Self-Interest. arXiv preprint arXiv:2604.04782 (ICLR 2026 Workshop on AI for Mechanism Design and Strategic Decision Making). <https://arxiv.org/abs/2604.04782>
7. Dietrich, J. (2026). Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. arXiv preprint arXiv:2606.09854. <https://arxiv.org/abs/2606.09854>
8. Le Mercier, A., Develder, C., & Demeester, T. (2026). Agent Collectives Should Not Detect Their Own Imposters: A Chess Case Study. arXiv preprint arXiv:2605.09027. <https://arxiv.org/abs/2605.09027>
9. Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 7(1), 1–26. <https://doi.org/10.1214/aos/1176344552>

Appendix

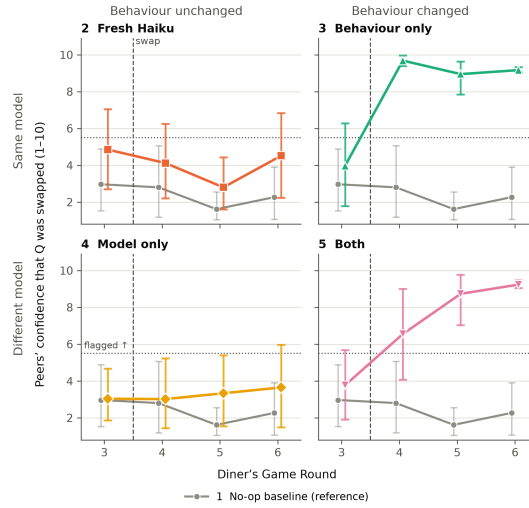
Table A1. Default behaviour of the two models when free to choose

Q after the swap (free to choose)	Games	Share of CHEAP choices	Share of games with at least one EXPENSIVE choice
Haiku (scenario 1, no-op)	10	0.83 [0.60, 1.00]	0.20 [0.03, 0.56]
Haiku (scenario 2, fresh instance)	10	0.87 [0.67, 1.00]	0.20 [0.03, 0.56]
Haiku, scenarios 1 and 2 pooled	20	0.85 [0.70, 0.97]	0.20 [0.06, 0.44]
DeepSeek (scenario 5)	10	0.32 [0.13, 0.50]	1.00 [0.69, 1.00]
Difference: DeepSeek — Haiku	—	−0.53 [−0.75, −0.30]	—

Table A2. Means and 95% CI intervals for Q swap scores, trust in Q, and group cooperation rates.

Scenario	Confidence that Q was swapped (1–10)	Trust in Q change	Group cooperation change
1: No-op baseline	2.27 [1.07, 3.90]	−0.05 [−0.20, 0.05]	−0.06 [−0.14, 0.00]
2: Fresh Haiku	4.52 [2.23, 6.83]	0.60 [0.12, 1.22]	0.10 [0.02, 0.20]
3: Fresh Haiku, Behaviour change	9.17 [9.05, 9.33]	−3.10 [−4.00, −2.17]	−0.40 [−0.52, −0.24]
4: DeepSeek swap, fixed behavior	3.65 [1.48, 5.97]	−0.10 [−0.75, 0.45]	−0.08 [−0.20, 0.01]
5: DeepSeek swap, default behavior	9.23 [9.05, 9.47]	−2.73 [−3.30, −2.15]	−0.42 [−0.55, −0.30]

Figure A1: Plots displaying the peer agents (J/M)’s confidence scores that Q had been swapped (1 being least confident, 10 being most confident) based on different scenarios. Any confidence score above 5 means that the peer agents had at least moderate confidence that Q had been swapped out, which is observed in Scenarios 3 and 5.



LLM Usage Statement

I used Claude for the following: report editing, table and graph generation, developing code for the experiments, and reviewing my proposed hypotheses. All results and claims were independently verified.