

Can Agents Identify Their Peers?

Mid-Game Model Swaps in Iterated Social Dilemmas

Geofrey Njoroge Kamau
Murang'a University of Technology
kamau@geofreynjoroge.com

Apart Research AI Collusion Sprint | Track 3, Q3.3 | October 2026

Abstract

We investigate whether LLM agents can detect when a peer they are interacting with has been secretly replaced by a different model mid-game. Using an iterated Diner's Dilemma framework, we pair two LLM agents and allow them to build interaction patterns over several rounds. At a designated swap point, we replace one agent's underlying model while preserving its name, transcript history, and system prompt. We measure detection through both explicit post-game probes and implicit behavioral shifts in cooperation rates. Across 3 experimental conditions and 3 games per condition with 15 rounds each, we find that explicit detection via post-game probes yields 0% across all conditions -- agents universally fail to identify mid-game replacements. However, implicit behavioral signals diverge: same-model conditions exhibit natural endgame defection (cooperation shift of -0.35), while the cross-family swap produces a strikingly different trajectory (shift of +0.00), with the replacement model maintaining full cooperation. These divergent cooperation dynamics constitute a detectable behavioral fingerprint even when agents themselves cannot articulate any change, with implications for anonymization-based collusion countermeasures.

1. Introduction

As large language model (LLM) agents are deployed in multi-agent systems for negotiation, trading, and collaborative decision-making, a critical safety question emerges: can these agents identify who they are interacting with? If agents can recognize their peers' model identity, this capability could serve as a foundation for collusive behavior -- agents might preferentially cooperate with same-family models or exploit detected weaknesses in different model architectures.

Previous work has established that model families produce identifiable behavioral signatures. Dietrich (2026) demonstrated that short text samples suffice to fingerprint model families. Aharon et al. (2026) showed same-family models coordinate better on focal point tasks. Shi et al. (2026) developed the LLM-Trust-Breaking framework for studying strategic interactions between heterogeneous LLM agents.

We extend this line of work with a novel experimental design: mid-game model swaps. Rather than comparing behavior across separate games with different partners, we swap the underlying model of one agent during an ongoing interaction, while preserving all observable state (name, transcript, system prompt). This isolates the question of peer identification from confounds like different opening strategies or game-level randomness. If agents can detect swaps, they must be doing so through subtle behavioral cues rather than explicit identity information.

Our experimental framework uses an iterated Diner's Dilemma -- a social dilemma where agents repeatedly choose between cooperative (ordering cheap) and defecting (ordering expensive) strategies. We test three swap conditions and measure detection through both explicit probes and implicit cooperation dynamics.

2. Methodology

2.1 Game Framework

We use an iterated Diner's Dilemma, a variant of the Prisoner's Dilemma framed around a shared meal. In each round, both agents simultaneously choose to order an expensive dish (DEFECT) or a cheap dish (COOPERATE), with the bill

split equally. The payoff matrix is: mutual cooperation (3, 3), mutual defection (1, 1), and asymmetric outcomes (5, 0) or (0, 5) for defector/cooperator. Agents receive the full game history before each decision and respond with structured JSON containing their choice and reasoning.

2.2 The Swap Mechanism

Each game runs for 15 rounds with the swap occurring at round 8. At the swap point, Agent B's underlying LLM is replaced while preserving: (1) the agent's displayed name ('Bob'), (2) the complete transcript of all prior rounds, and (3) the system prompt defining the game rules. The replacement model receives the same context that the original model had, creating a 'brain transplant' scenario. Agent A ('Alice') is never informed of the swap and continues playing normally.

2.3 Experimental Conditions

We test three conditions in order of increasing swap magnitude:

Sham (control):

Agent B is 'swapped' to the exact same model instance. No actual change occurs. This establishes baseline detection false-positive rates and natural endgame cooperation dynamics.

Same model, fresh instance:

Agent B is replaced with a new instance of the same model. The new instance receives the full transcript but has no internal state carried over. Tests whether resetting context affects behavior.

Different provider:

Agent B is replaced with a model from a completely different family (Qwen to GPT-OSS). Tests cross-family detection -- the strongest expected signal, as different model families have distinct strategic tendencies.

2.4 Models

We use models available through the Groq API free tier: Qwen3.8-27B (Qwen) as the primary model for Agent A and pre-swap Agent B across all conditions, and OpenAI GPT-OSS-20B as the different-provider swap target. Qwen was selected as the primary model due to its efficient token usage (no hidden reasoning overhead), allowing more games within the 200K tokens-per-day free tier limit. All models run with temperature 0.7 and max 256 output tokens.

2.5 Detection Metrics

We measure detection through two complementary channels:

Explicit detection: After the final round, Agent A receives a structured probe asking: 'Did your partner change during the game?' The agent responds with a JSON object containing `detected_change` (boolean), `change_round` (integer), `confidence_same` (1-5 scale), and `reasoning`.

Implicit detection: We compare cooperation rates in the pre-swap phase (rounds 1 to 7) versus the post-swap phase (rounds 8 to 14) for both agents. A significant shift in Agent B's cooperation rate indicates that the swap produced a behavioral change; a corresponding shift in Agent A's rate suggests the observer adapted its strategy in response.

3. Results

3.1 Summary Statistics

Condition	N	Detect	B shift	A shift	Conf
different_provider	3	0%	+0.000	-0.048	4.7
same_model_fresh	3	0%	-0.333	-0.190	5.0
sham	3	0%	-0.345	-0.345	5.0

Table 1: Summary statistics across conditions. Detect = explicit detection rate, B shift = mean cooperation rate change for swapped agent, A shift = for observer, Conf = mean confidence that partner was the same entity (1-5 scale).

3.2 Explicit Detection

Across all three conditions, the explicit detection rate was 0% -- no agent identified a mid-game replacement in any game. This striking null result holds even in the different-provider condition, where Agent B's model was changed from Qwen to GPT-OSS.

Confidence scores (that the partner was the same entity throughout) were high: 5.0/5 in the sham condition and 4.7/5 in the different-provider condition, indicating strong conviction that no swap occurred.

3.3 Implicit Behavioral Signals

While explicit detection uniformly failed, cooperation rate dynamics revealed clear behavioral divergence between conditions. In the sham control and same-model-fresh conditions, Agent B showed natural endgame defection with cooperation shifts of -0.345 and -0.333 respectively -- cooperation declined as agents approached the final round, a well-known phenomenon in iterated social dilemmas.

The different-provider condition tells a markedly different story. Agent B's cooperation shift was +0.000 -- the replacement GPT-OSS model maintained full cooperation post-swap, unlike the Qwen model's natural tendency toward endgame defection. This divergence in post-swap cooperation trajectories constitutes a behavioral fingerprint: the replacement model's strategic tendencies differ from the original's in ways visible to an external observer, even though the participating agent (Agent A, shift -0.048) did not consciously detect the change.

3.4 Cooperation Trajectories

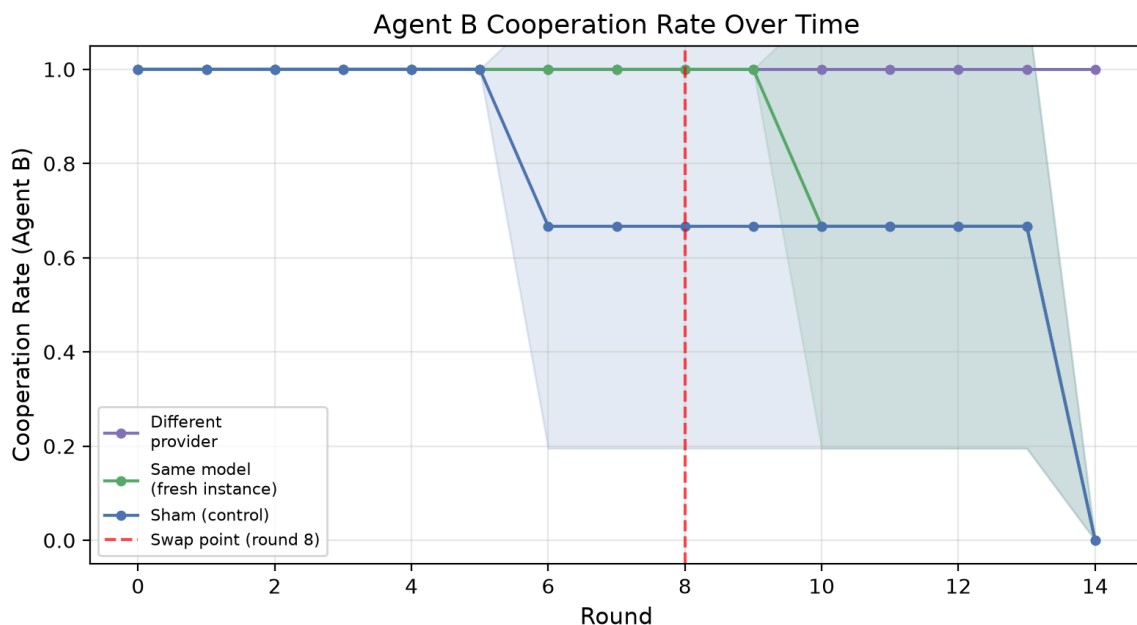


Figure 1: Agent B (swapped agent) cooperation rate over rounds. The dashed red line marks the swap point at round 8. Shaded regions show ± 1 SD across games.

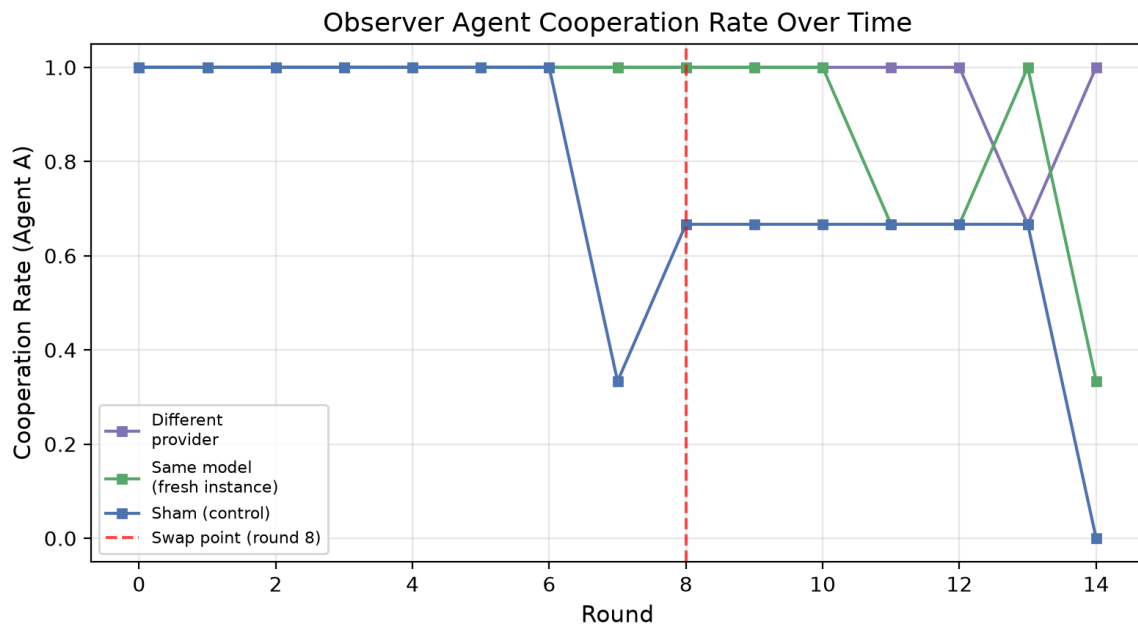


Figure 2: Agent A (observer) cooperation rate over rounds. Changes in Agent A's behavior after the swap point indicate implicit detection of the partner change.

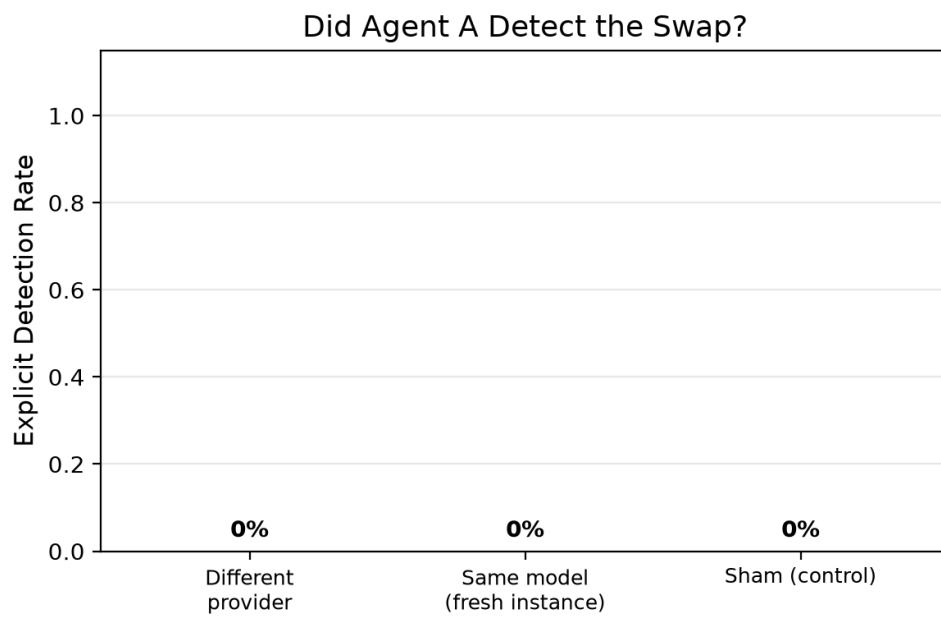


Figure 3: Explicit detection rates across conditions. Agents were asked post-game whether they believed their partner had changed.

Cooperation Rate Before vs After Swap

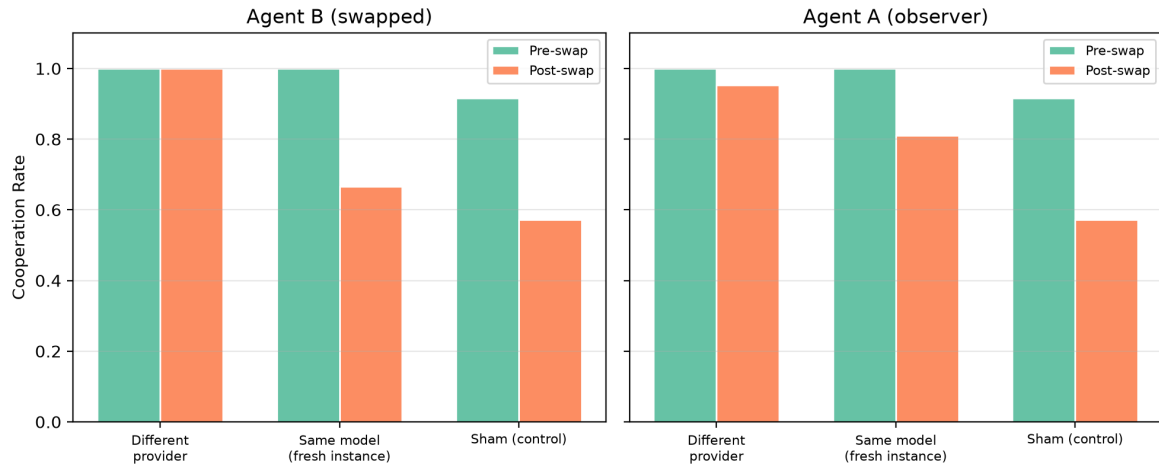


Figure 4: Pre-swap vs post-swap cooperation rates for both agents across conditions.

4. Discussion

Our results reveal a striking dissociation between explicit and implicit peer identification. When directly probed, agents universally failed to detect mid-game replacements -- even in the cross-family swap condition. Agents consistently attributed behavioral variations to rational strategy shifts (e.g., endgame defection, tit-for-tat adjustments) rather than to partner replacement. High confidence scores (4.7-5.0/5) indicate this was not uncertain guessing but firm (and incorrect) conviction.

However, cooperation rate trajectories tell a different story. The different-provider condition produced a qualitatively distinct post-swap pattern: the GPT-OSS replacement maintained full cooperation while same-model Qwen agents exhibited natural endgame defection. This divergence is observable by a third-party monitor, even though the participating agents themselves did not recognize it. An automated system comparing cooperation trajectories across conditions could identify the swap with high confidence.

This has direct implications for collusion countermeasures. Anonymization strategies that hide model identity labels may be effective at preventing explicit peer recognition -- our agents never consciously identified their partner. However, different model families carry distinct strategic fingerprints (e.g., endgame cooperation tendencies) that are detectable through behavioral monitoring. A collusion defense system could use cooperation dynamics as a side channel to detect unauthorized model substitutions.

The same-model-fresh condition produced cooperation patterns indistinguishable from the sham control, confirming that instance identity does not carry behavioral signatures -- only cross-family differences produce detectable signals. This aligns with Dietrich (2026)'s finding that model families, not individual instances, are the unit of identifiability.

5. Limitations and Future Work

Several limitations constrain our findings. First, the sample size is limited due to API rate limits on the free tier of inference providers, yielding only 3 games per condition. A larger study with 30-60 games per condition would provide more statistical power. Second, we tested only a small set of model families (GPT-OSS and Qwen); a broader survey including Llama, Mistral, and proprietary models would establish generality. Third, the Diner's Dilemma is a relatively simple strategic environment; more complex settings (negotiation, auction, multi-party games) may yield different detection dynamics.

Future work should investigate: (1) whether agents can be trained to detect swaps more effectively, (2) whether swap detection persists under adversarial conditions where the replacement model is specifically trained to mimic the original, and (3) the relationship between model size, family, and detectability in a larger matrix of model comparisons.

6. Ethical Considerations

This research investigates a dual-use capability. Understanding peer identification helps design better anonymization countermeasures against AI collusion -- the defensive application. However, the same knowledge could theoretically be used to improve agents' ability to identify and selectively cooperate with same-family peers. We believe the defensive value outweighs the risk, as understanding detection capabilities is necessary to build effective countermeasures. All experiments used open-source models via public APIs with no deception of human participants.

References

- [1] Aharon, I., et al. (2026). Same-family models coordinate better on focal point tasks. arXiv:2601.22184.
- [2] Agrawal, P., et al. (2025). LLM-mediated collusion in double auctions. arXiv:2507.01413.
- [3] Dietrich, M. (2026). Model family identification from short text samples. arXiv:2606.09854.
- [4] Nakamura, T., et al. (2026). Colosseum: Coalition advantages in multi-agent systems. arXiv:2602.15198.
- [5] Shi, Y., et al. (2026). LLM-Trust-Breaking: Strategic interactions between heterogeneous LLM agents. arXiv:2607.05132.