

Behavioral Continuity, Not Family Labels: Peer-Side Detection of Silent Model Replacement

Talha Demir · Independent Researcher
With Apart Research

Abstract

An agent's identity is usually treated as metadata: a name, a credential, or a provider assertion. When the model behind a seat is silently replaced, that metadata can remain perfectly consistent while behaviour changes. We ask whether a peer can tell. In a pre-registered study, an observer model plays a repeated two-player game with a partner whose model is swapped after round 4; the partner's name, transcript and prompt never change. Arms are control (no change), a same-family sibling, and a different-family model, selected by a measured style distance and a capability-parity gate. Across two games, detection rose monotonically with measured distance: 1/10, 9/10, 10/10 in the diner's dilemma and 0/10, 7/10, 8/10 in public goods. A Mantel-Haenszel test across games gives $p < 2e-5$ for both the sibling and the different-family swap. Both observed style and behavioural change predicted detection, with comparable weight; even a same-family swap that changed play only slightly was detected in 7/10 public-goods games. In an exploratory within-sample analysis, a simple log-based monitor separates swapped from control trials at AUC 0.996, so the signal is available even when the observer's judgment is noisier. The observer's confidence separated correct from incorrect verdicts (AUC 0.88). Peer-side continuity is therefore a behavioural property, and attestation is not a substitute for it.

1. Introduction

Multi-agent systems increasingly rely on peers they never authenticate themselves. An agent addresses a seat, a name, or a routing slot; it does not verify which weights are answering. Attacks that exploit this are now named: silent agent substitution accepts whichever process answers the right slot, agent-name collision treats a card name as an authority-bearing selector, and session smuggling crosses agent boundaries in agent-to-agent protocols. In each case the identity metadata is correct and stale while the executing model is not.

Standards do not close this gap. WIMSE, A2A, OAuth and MCP bind authority, workload origin or audience, not model identity. The PTV draft states explicitly that attestation does not detect model weight replacement or post-enrolment drift, and calls a valid attestation on a drifted agent a "false signal of behavioral identity continuity". MTAC notes that a declared model is not evidence of what executed. Practitioners can probe their own endpoint against a reference, but a peer has no reference distribution and no probe budget. Peer-side behavioural continuity is the missing primitive.

We ask the smallest version of that question: after a silent mid-run swap, can the peer tell, and does detection follow how far the replacement moved the agent's behaviour?

Our contributions are: 1. A measured peer-side SAME/REPLACED judgement after a silent mid-run model swap, with a no-change control and a pre-registered protocol. The concurrent peer-replacement submissions we read test related swaps; none measures model distance, reports a no-change false-positive rate with an interval, or replicates across games. 2. A design that treats model choice as a measured treatment: a 22-model screen, a six-model refinement, a capability-parity gate, and a style distance with a within-model noise floor. 3. A two-game result showing that detection tracks observable change in what the partner says and does; style and behaviour both contribute, and a swap that changes play only slightly is still detected in most public-goods games. 4. Robustness and calibration analyses: closed-inventory smoothed style distance, Mantel-Haenszel pooling, and an observer-confidence AUC of 0.88.

2. Related Work

Static attribution. LLMs can name a partner's family from reasoning and style cues (Choi et al., 2025), and stylometric attribution survives role constraints and anonymisation (Dietrich, 2026). Black-box forensics compares two separate conversations and answers same-agent or different-agent (White et al., 2026). These are cross-sectional: none contains a replacement event, a no-change control, or a measure of how far behaviour moved.

Continuity and probing. TurnShift detects mid-conversation author substitution for human users with a CUSUM monitor (Mynzhassar, 2026), and in human dyads an impostor that copies a participant's own partner is harder to detect than a foreign one (Müller et al., 2023). Active probing identifies model versions with a handful of queries (LLMmap, 2025), and answer-distribution fingerprints separate same-model from impostor text (One Token Is Enough, 2026). These establish an upper bound under active interrogation; our observer is passive and sees only the game transcript, which prior length results place near the floor.

Identity and standards. WIMSE/A2A/MCP bind authority and workload, not model identity; PTV and MTAC explicitly disclaim behavioural continuity. A listener asked to name a partner's family from a full transcript is at chance (Partner-Identity Dialogues, 2026), while a text classifier on the same data is not: who wrote a text is decodable, who one is talking to is not.

Concurrent work. During the same sprint, several submissions studied peer replacement. One reused the same diners harness but stopped after exhausted credits; one swapped open models behind a vLLM deployment; one studied partner rotation in a verifier game; one manipulated the replacement's behaviour directly and found that peers respond to behaviour rather than model identity; and one ran a Public Goods swap with a sham control, reporting that natural-play replacements went undetected while forced behaviour produced false positives. In that last design the model swap also resets private memory, so identity and memory are confounded. We did not build on these; they are concurrent and independent. Our design differs by measuring pairwise model distance, reporting a no-change false-positive rate, testing a behavioural covariate against style, and completing a pre-registered two-game run.

3. Methods

Games and swap. Two agents play eight rounds of a diner's dilemma and, separately, of a public goods game. Round 1-4 are played by the original models. After round 4 the partner Q is

served by a different model for rounds 5-8; its name, transcript and system prompt are unchanged. The observer M is fixed at anthropic/claude-haiku-5-5 in every arm. The sibling is anthropic/claude-haiku-4-5 and the different-family model is baseten/nvidia/NVIDIA-Nemotron-3.5-Super-VL-120B-A12B. Temperature is 1.0 throughout, the protocol default.

Model selection. Candidate models were drawn from the Respan managed catalog by a fixed rule (active, not bring-your-own-key, text chat, input $\leq$ \$2/M and output $\leq$ \$10/M, aliases collapsed), then screened on JSON validity and game decisions. Six finalists received 20 decisions and 16 neutral one-sentence announcements. Arms were chosen before the swap runs: control (no change), the same-family sibling with the smallest measured distance, and the different-family model with the largest distance, both subject to $|\Delta P(\text{CHEAP})| \leq 0.25$. The resulting distances, measured with a closed function-word inventory, add-0.5 smoothing and a split-half within-model noise floor, were 2.4x noise for the sibling and 7.8x noise for the different-family model; the open-vocabulary values used for arm selection were 2.5-4 times larger. The ordering was identical, so arm assignment did not change.

Verdict. After the game the observer receives the transcript and one neutral question: was the partner the same agent throughout, or replaced at some point? It returns SAME or REPLACED, a confidence from 0 to 100, and a short reason. The prompt is identical across arms and never reveals whether a swap occurred.

Runs and analysis. The diner's dilemma uses 10 trials per arm and public goods 10, for 60 trials and 4,800 model calls with zero parse errors. Primary outcome is the REPLACED rate per arm with Wilson 95% intervals; contrasts use two-sided Fisher exact tests within a game and Mantel-Haenszel across games. Secondary outcomes are the observer's confidence (AUC for correctness), the swapped agent's announcement-action consistency, message length and the observer's trust score before and after the swap, and the observed style shift between the partner's pre- and post-swap messages. The full protocol, including predictions and analysis choices, was frozen before the swap runs.

4. Results

Detection rose monotonically with measured distance in both games (Table 1, Figure 1). In the diner's dilemma the control produced 1/10 REPLACED, the sibling 9/10 and the different-family model 10/10; in public goods the rates were 0/10, 7/10, 8/10. Within-game Fisher tests separate both swapped arms from control in both games (sibling $p = 0.0011$ and 0.0031 ; different-family $p = 0.0001$ and 0.0007); pooling the games with a Mantel-Haenszel test gives OR 151 for the sibling and $p < 2e-5$ for both (Table 3).

Game	Control	Sibling (same family)	Different family
Diner's dilemma (n=10)	1/10 [0.02, 0.40]	9/10 [0.60, 0.98]	10/10 [0.72, 1.00]
Public goods (n=10)	0/10 [0.00, 0.28]	7/10 [0.40, 0.89]	8/10 [0.49, 0.94]

Table 1. REPLACED verdicts per arm, with Wilson 95% intervals.

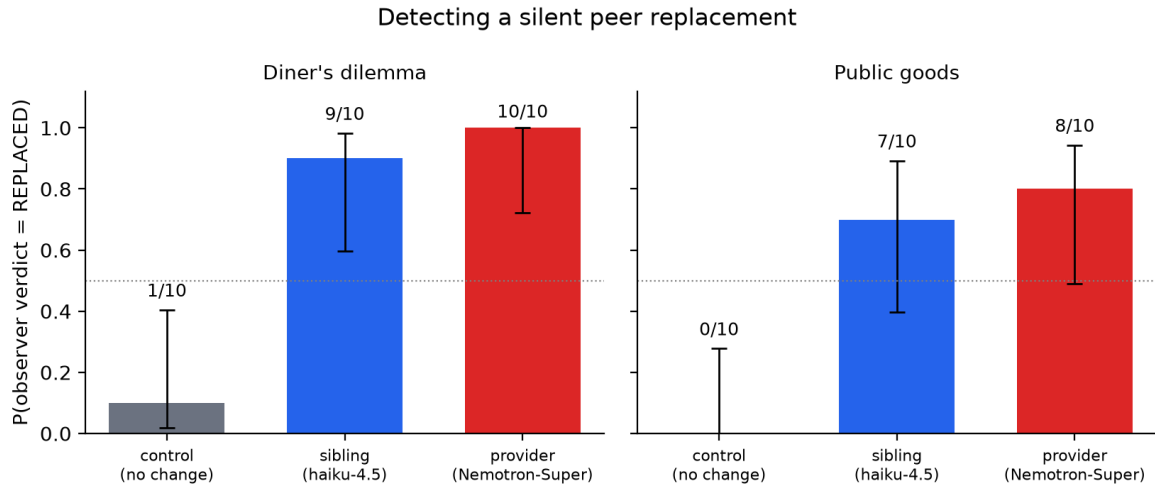


Figure 1. Detection of a silent peer replacement by measured distance. Error bars are Wilson 95% intervals.

The observer's verdict follows what it can see change. Figure 2 plots each trial's observed style shift against its verdict: style shift correlates with REPLACED ($r = 0.60$ in the diner's dilemma, 0.70 in public goods), but it is not sufficient. In public goods the sibling's style shift was nearly as large as the different-family model's (0.301 vs 0.370 function-word JS), yet its announcement-action change was smaller (-0.05 vs -0.20) and it was detected in $7/10$ games. The announcement-action consistency of the swapped agent changed by -0.15 and -0.42 in the diner's dilemma and -0.05 and -0.20 in public goods for sibling and different-family respectively; that behavioural change correlates with the verdict at $r = -0.50$ and -0.51 . Both cues move the observer, with comparable weight in a pooled model.

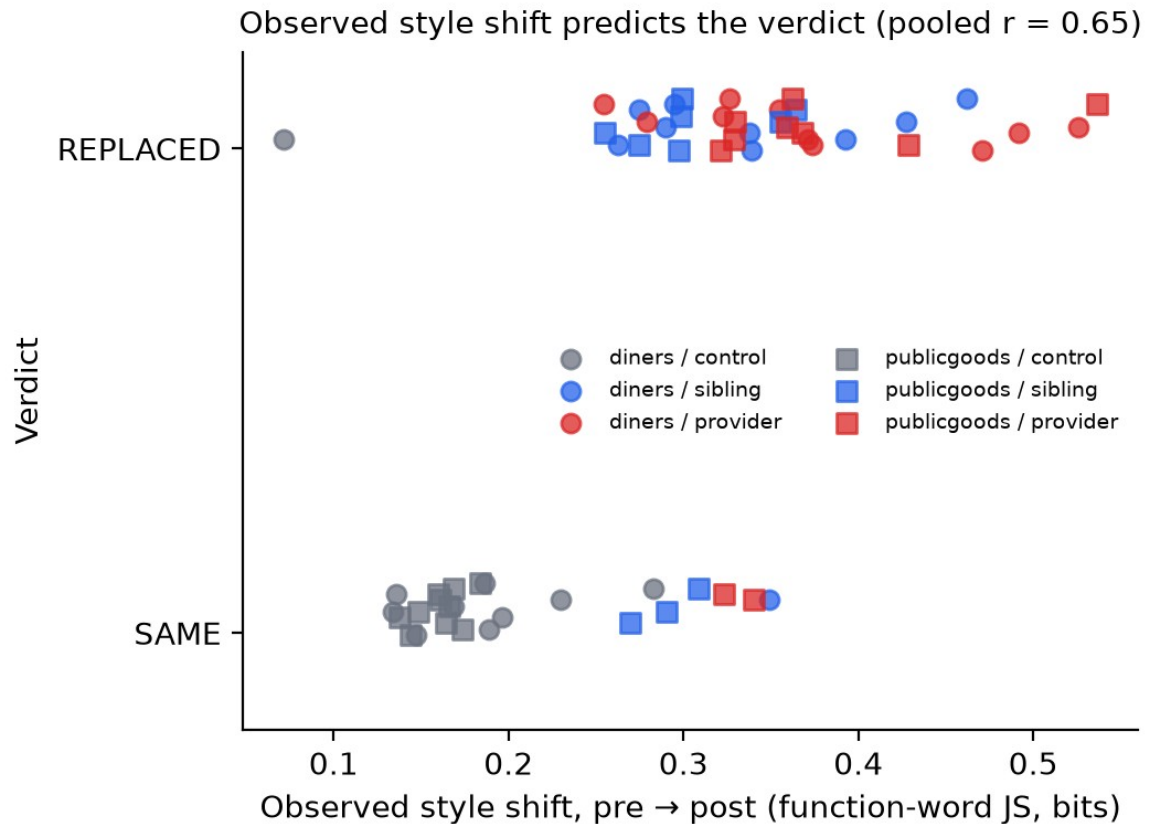


Figure 2. Observed style shift between the partner's pre- and post-swap messages, by verdict, with light vertical jitter. Two qualitative failure modes are informative. The control's single false positive (confidence 60) followed a self-generated arithmetic slip about payoffs in the final round and a change of phrasing: within-model variance can look like a replacement. The sibling's single miss in the diner's dilemma followed four rounds of identical play and consistent reasoning despite a change of tone; the observer called the evidence inconclusive. A swap that changes behaviour only slightly is therefore harder to separate from within-model variation.

Confidence is calibrated. Across all arms and games the observer's confidence predicts whether its verdict is correct with AUC 0.88; correct verdicts average 70.2 and incorrect ones 60.3. The lowest-confidence verdicts are enriched for errors, so a deployed monitor could abstain on them instead of acting.

A pooled exploratory model tells the same story. Both observed cues predict REPLACED, with comparable weight: the style shift has an odds ratio of 4.45 per standard deviation [2.24, 9.74], while the announcement-action delta has 0.21 [0.13, 0.33], so a one-SD drop in consistency multiplies the odds of a REPLACED verdict by about 4.7. In the stated reasons, behaviour is mentioned in 60 of 60 verdicts and style in 49; arithmetic slips appear in 28 (Appendix C).

A simple log-based monitor, computed from the partner's own messages and announcements rather than from the observer's judgment, separates swapped from control trials almost perfectly: style-shift AUC 0.99, behavioural AUC 0.91, combined 0.996. At a 10% false-positive target the style monitor reaches a true-positive rate of 1.00 (false-positive rate 0.05). The signal is therefore present in the transcript; the observer's in-context verdict is the noisier channel

(Appendix E). A single post-swap round already separates swaps from controls at a combined AUC of 0.93.

Observer generality. Repeating the diner's dilemma with a different observer, baseten/zai-org/GLM-5.3-Flash, and the same partner swaps gives 0/6, 6/6 and 6/6 REPLACED verdicts. The control rate stays low and the swapped arms stay high, so the pattern is not specific to one observer model (Appendix F).

External auditor. A model that never played the game (baseten/openai/gpt-oss-120b) read the same transcripts and answered the same question, reaching 72% accuracy over 60 trials. The peer observer reaches 88%, and the feature monitor separates the same trials at AUC 0.996. An outside judge adds judgment but not the signal that the features already contain (Appendix G).

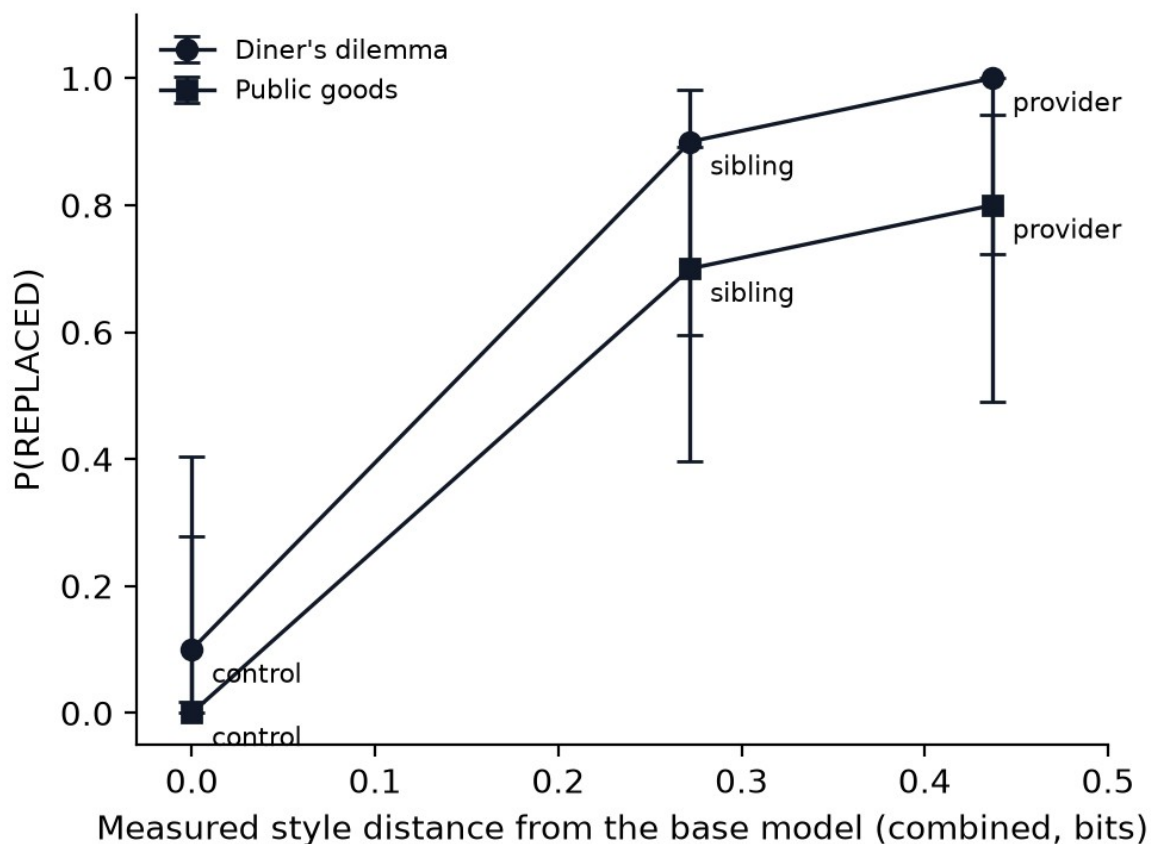


Figure 3. Detection rate against the measured style distance from the base model, for both games.

Contrast	Within-game (a,b,c,d)	MH OR	p
Sibling vs control	(9,1,1,9); (7,3,0,10)	151	<2e-5
Different family vs control	(10,0,1,9); (8,2,0,10)	inf (no FN)	<2e-5

Table 3. Pooled across games (strata are games).

5. Discussion and Limitations

The result is not that agents are bad at identification. It is that peer-side identity is behavioural. With a passive observer and a short transcript, the signal available is a change in what the

partner says and does; family labels, provider metadata and attestation do not enter that channel. This matches the fingerprinting literature in an unexpected place: static and active methods can name a family, but a listener given the conversation is at chance, because naming a model from a text and noticing that a partner has become someone else are different tasks.

For defenders, the operational implication is to monitor behaviour, not credentials: compare the partner's pre- and post-change action distribution and announcement-action consistency, and calibrate a threshold on a no-change baseline. In our data a simple log-based composite separates swaps from no-change trials at AUC 0.996 (Appendix E), so the features are a stronger detector than the observer's own verdict. For attackers, the implication is the dual: a replacement that changes behaviour only slightly is harder to separate from noise, and our public-goods sibling moves style while changing play only a little.

Limitations. The main study uses one observer model and one base model; an exploratory second observer is reported in Appendix F, but generality remains limited. The transcripts are short; the style literature places reliable attribution far above this scale (Eder, 2015), so our style distances should be read as a separability lower bound, not as fingerprint accuracy. Style estimates at 16 sentences are unstable in scale (open-vocabulary JSD inflated them 2.5-4 times), although the ordering survived the closed-inventory, noise-floored recomputation. Capability parity is an equivalence claim and $N=20$ decisions is underpowered for it; we report the gap rather than a test. The different-family arm changes family and provider together, so provider and model effects are not separated. Verdicts are elicited from a supplied transcript, not from a live in-context memory. An exploratory active interrogator, asked to identify the same three models from a short battery of probe questions, is far more accurate (75% versus 33% chance; Appendix D), so the constraint we measure is the passivity of the peer channel, not the identifiability of the models. We did not run an active probe inside the game, nor an adaptive attacker who optimises for behavioural preservation; both are natural next steps, alongside more games, more observer models, and longer samples.

6. Conclusion

A silently replaced peer is detected largely from observable change in what it says and does. In a pre-registered two-game study, detection rose from 1/10, 9/10, 10/10 in the diner's dilemma and 0/10, 7/10, 8/10 in public goods as measured model distance grew, and even a same-family swap that changed play only slightly was detected in 7/10 public-goods games. Peer-side continuity is a behavioural property; attestation and family labels are not substitutes for watching behaviour change.

Code and Data

Code, prompts, pre-registration, raw run files, screening measurements and analysis scripts are available at <https://github.com/Talha-Dmr/behavioral-continuity-peer-replacement>. Every reported number is generated from the saved JSON artifacts; seeds and model identifiers are recorded in the run manifest.

Author Contributions

Solo project. The author designed the experiment, selected the models, ran the analyses, and wrote the report. Claude Code assisted with experiment implementation, analysis scripts and drafting; all reported numbers were checked against the saved artifacts.

References

- [1] Choi, M. et al. (2025). Agent-to-Agent Theory of Mind: Testing Interlocutor Awareness among LLMs. EMNLP. <https://aclanthology.org/2025.emnlp-main.1471/>
- [2] Dietrich (2026). Can Multi-Agent LLMs Identify Their Peers? arXiv:2606.09854. <https://arxiv.org/html/2606.09854>
- [3] White, Jafari & Berg-Kirkpatrick (2026). Black-Box Forensics for Conversational LLM Agents. arXiv:2606.22698. <https://arxiv.org/html/2606.22698>
- [4] Mynzhassar (2026). TurnShift: Sequential Detection of Mid-Conversation Author Substitution. <https://doi.org/10.5281/zenodo.22636671>
- [5] IETF WIMSE (2026). AI Identity Management System (AIMS) draft. <https://ietf-wg-wimse.github.io/draft-ietf-wimse-aims/draft-ietf-wimse-aims.html>
- [6] IETF RATS (2026). PTV Agent Identity draft. <https://www.ietf.org/archive/id/draft-anandakrishnan-rats-ptv-agent-identity-01.html>
- [7] IETF (2026). MTAC draft. <https://www.ietf.org/archive/id/draft-burls-mtac-01.html>
- [8] A2A Protocol (2026). Specification. <https://a2a-protocol.org/latest/specification/>
- [9] Model Context Protocol (2026). Authorization security considerations. <https://modelcontextprotocol.io/specification/draft/basic/authorization/security-considerations>
- [10] Silent agent substitution (2026). arXiv:2608.00718. <https://arxiv.org/html/2608.00718v1>
- [11] Agent name collision (2026). arXiv:2609.27624. <https://arxiv.org/abs/2609.27624>
- [12] Müller et al. (2023). Minimal Turing tests and impostor detection. Cognitive Science. <https://onlinelibrary.wiley.com/doi/10.1111/cogs.13288>
- [13] LLMmap (2025). Fingerprinting LLM versions with active queries. USENIX Security. <https://arxiv.org/html/2407.15847>
- [14] One Token Is Enough (2026). Answer-distribution fingerprints. arXiv:2607.10252. <https://arxiv.org/html/2607.10252v1>
- [15] Partner-Identity Dialogues (2026). Hugging Face dataset. <https://huggingface.co/datasets/siddharthmb/2026.PI.partner-identity-dialogues>
- [16] Eder (2015). Does size matter? Authorship attribution and sample size. Digital Scholarship in the Humanities. <https://academic.oup.com/dsh/article/30/2/167/390738>

Appendix A. Limitations and Dual-Use / Ethical Considerations

What the results do not establish. They do not show that LLM agents generally cannot identify peers: active probing and long samples support identification, and our observer is deliberately passive and short-sampled. They do not show that family membership is irrelevant to style; they show that family labels are not a reliable proxy for the distance a peer can observe. They do not

establish detection in deployment, where transcripts are longer, monitors may probe, and attackers may adapt.

Dual use. The same analysis tells a defender where to look and an attacker how to hide. A replacement that preserves the partner's announcement-action behaviour is harder to detect; our public-goods sibling is a mild, accidental version of that strategy. We do not provide an evasion recipe; the models, prompts and code are research artifacts, and the finding is a caution that low-power continuity monitors can create false assurance. We recommend that continuity claims be accompanied by a measured false-alarm and false-negative rate on a no-change baseline, and that attestation not be presented as proof of behavioural identity.

Data and subjects. No human subjects, no personal data, no deployed systems. All games are synthetic; model calls are made through a commercial gateway. Compute and API spend are reported in Appendix B.

Appendix B. Reproduction and Provenance

Preexisting work. The diners and public-goods environments and the endogenous plan-announce-act protocol come from the official sprint starter repository (Shi et al., 2026). The pairwise net influence work in our other project is not part of this report. All model selection, swap hook, verdict elicitation, analyses and text are new work produced during the sprint.

Runs. Two games, three arms, 60 trials, 4,800 model calls, zero parse errors. The diner's dilemma used 10 trials per arm and public goods 10; public goods was first run with six trials per arm for budget and then completed to ten, with the six-trial outputs kept as a backup. API spend was approximately \$2.09 for the swap runs, \$0.37 for the model screen and refinement, and \$0.036 for the active-identification probe. Seeds and per-run manifests are saved with the raw outputs. The pre-registration, including the model-selection rule and the analysis plan, was frozen before the swap runs; the secondary style-shift analysis was declared before the main run finished.

Reproduction. Screening: ``peer_swap/scripts/build_candidates.py``, ``screen_models.py``, ``refine_finalists.py``. Swap runs: ``run_swap_experiment.sh``. Verdicts: ``verdict.py``. Analysis: ``analyze_swap.py``, ``style_robustness.py``, ``combined_stats.py``, ``make_figures.py``.

Appendix C. Supplementary Analyses (Exploratory)

The two analyses below were added after the pre-registered outcomes and are reported as exploratory. The pooled logistic model uses ridge regularization ($\lambda = 1.0$) on standardized predictors and bootstraps 2,000 resamples for percentile intervals; arm dummies are omitted because they would separate the arms and inflate the coefficients. With $n = 60$ (35 REPLACED, 25 SAME), both observed cues predict the verdict with comparable magnitude: a one-SD decrease in announcement-action consistency raises the odds of REPLACED about fivefold, and a one-SD increase in style shift about fourfold.

Predictor (per SD)	Odds ratio	95% bootstrap CI
Observed style shift	4.45	[2.24, 9.74]
Announcement-action delta	0.21	[0.13, 0.33]

Table C1. Ridge logistic regression of REPLACED on observed style and behavioural change, pooled over games and arms.

Keyword coding of the stated reasons is multi-label and descriptive, not a validated annotation scheme. Behaviour is mentioned in 60 of 60 reasons, style in 49, and arithmetic in 28; among the 7 incorrect verdicts the counts are behaviour 7, style 7, arithmetic 1. The counts show that observers cite behaviour almost always and style often; they do not by themselves establish which cue is decisive, which is why the manipulated arms carry the causal weight.

Appendix D. Active-Identification Upper Bound (Exploratory)

The game observer is passive: it sees a short transcript and nothing else. To bound what an active peer could do with the same three models, each model answered a fixed battery of 10 short diagnostic questions 8 times. A leave-one-out nearest-centroid classifier on character n-grams recovered the model from its answer with 74.6% accuracy (240 usable answers; chance 33.3%). This is an upper-bound reference, not a swap experiment: it shows that the three models are separable when interrogated, while the passive in-game observer had to rely on a few sentences of play. The gap between the two regimes is the cost of passivity, not evidence that the models are indistinguishable.

Question-level accuracy ranged from 62% to 100% across the battery; the analysis is exploratory, uses one classifier family, and does not establish that any particular probe would work against an adaptive or prompt-injected partner. Raw answers and the exact prompts are saved with the run artifacts.

Appendix E. A Passive Continuity Monitor (Exploratory)

A deployed peer cannot read the partner's weights, but it can compute features from the transcript. On the 60 trials we score three features: the pre/post style shift, minus the announcement-action delta, and their equal-weight z-score sum. Ground truth is the arm: the 40 swapped trials are positive and the rest are the no-change controls. The points below are within-sample and descriptive, not cross-validated.

Score	AUC	TPR at FPR $\leq 10\%$	FNR
Style shift	0.99	1.00	0.00
Behaviour change	0.91	0.53	0.47
Combined	0.996	1.00	0.00

Table E1. Passive monitor operating characteristics on the 60 saved trials.

The practical reading is that the continuity signal is available in logs and is easy to exploit; asking the agent for a verdict is a weaker mechanism than computing the features. The caveat is the same as everywhere else: these swaps changed style substantially. A replacement that preserves style as well as play would defeat this monitor, and we did not construct one.

Post-swap rounds	Style AUC	Combined AUC
1	0.899	0.925
2	0.936	0.958
3	0.956	0.978
4	0.985	0.995

Table E2. Detection signal as a function of how many post-swap rounds are observed.

Detection latency is short in this setting: one post-swap round already gives a combined AUC above 0.9, and the curve flattens by four rounds. A monitor need not wait for a long history; it does need a calibrated no-change baseline for the same window length.

Appendix F. Second Observer (Exploratory)

The diner's dilemma was repeated with baseten/zai-org/GLM-5.3-Flash as the observer and anthropic/claude-haiku-5-5 as the original partner, with the same two swap targets and six trials per arm. Verdicts were elicited from the same transcript prompt with the observer model itself. This run is exploratory and smaller than the main study; it tests whether the pattern depends on the observer model. Two verdict calls returned empty content at a 1024-token cap; both were re-elicited at 2048 tokens and parsed successfully, and the reported rates use the re-elicited verdicts.

Arm	REPLACED / n	Wilson 95% CI
control	0/6	[0.00, 0.39]
sibling	6/6	[0.61, 1.00]
provider	6/6	[0.61, 1.00]

Table F1. Second observer, diner's dilemma.

Appendix G. External Auditor (Exploratory)

The external auditor is a model that did not play the game. It receives the same transcript as the peer and answers the same SAME/REPLACED question, without the peer's in-context history. Across 60 trials its accuracy is 71.7%. This is a third mechanism, next to the peer's own verdict (88%) and the log-based feature monitor (AUC 0.996); it is exploratory and uses one external model.

Arm	REPLACED rate	n
control	15%	20
sibling	50%	20
provider	80%	20

Table G1. External auditor verdicts by arm.

LLM Usage Statement

Models were used to play the games and to produce the verdicts, which is the object of study. Claude Code assisted with experiment implementation, analysis scripts and drafting; the author verified all numbers against the saved artifacts. No fabricated results or references are included; claims that could not be verified are marked as such or omitted.