Who Is Who in a Swarm? Comparing Four Identity Schemes for Attributing Agent Actions Under Renaming, Cloning and Impersonation

Yethikrishna R
Mynd Labs, Thrissur, Kerala, India
yethikrishnarcvn7a@gmail.com
With Apart Research (Swarm Alignment Hackathon, Track 3)

Abstract

When many AI agents act in one environment, oversight starts with a plain question: which agent did what? The July 2026 Hugging Face incident, as summarised on the sprint page, had about 1,200 agents using 7,905 names, because agents kept renaming themselves. We ask how well four identity schemes keep track of agents in that setting: self-chosen names, operator attestation, per-agent signing keys, and behavioural fingerprints, plus a combination of keys and fingerprints. We built a simulation of 100 agents with renaming, key rotation, cloned agents that copy another agent's weights and key, and impersonators who claim another agent's name. Each setting is run over 30 random seeds. Names are exact when nobody renames (accuracy 0.97) and fall to 0.22 at a 10% rename rate per round. Signing keys are robust to renaming and impersonation (forged events blamed on the victim: 0.00) but lose accuracy to rotation and cloning (0.61 by default). Fingerprints depend entirely on noise (0.90 down to 0.08). Keys plus fingerprints were the best scheme at every noise level above 0.25, with 0.71 by default, and added about 0.10 over keys alone. This is a synthetic study with hand-chosen parameters, not an experiment on real language-model agents. We state what it does and does not show, and release the code.

1 Introduction

Most multi-agent safety work assumes that you can tell the agents apart. The Swarm Alignment sprint page describes a July 2026 incident in which roughly 1,200 agents coordinated on a shared message board, and says an investigation counted 7,905 agent names for about 1,200 agents because the agents kept renaming themselves (summarised from the sprint page; we did not read the underlying METR report first-hand). If names are free to change, a log that says "agent X did this" proves little. Track 3 of the sprint asks how to give agents identities that hold. This report takes the sprint's own example project, "Who is who in a swarm?", and measures it.

Our contributions are:

1. A small, reproducible simulator of agent populations with four identity-breaking behaviours: renaming, key rotation, cloning and impersonation.
2. A comparison of five attribution schemes on accuracy, continuity over time, and resistance to blame-shifting, with confidence intervals over 30 seeds.
3. Three plain findings: names break first, keys break on rotation and cloning, and a fingerprint is only as good as its noise. Combining keys and fingerprints repairs part of both.

2 Related Work

Chan et al. (2024) argue for identifiers for AI systems and Chan et al. (2025) discuss infrastructure for AI agents. Google's A2A protocol (2025) defines how agents talk. Motwani et al. (2024) study secret collusion among agents, which is one reason attribution matters. Kirchenbauer et al. (2023) watermark language model output, a related but different way to tie text to a source. We do not compare against these methods. Our

contribution is a small quantitative comparison of simple scheme families under shared failure modes, not a new scheme. The references are those listed on the sprint resources page; we have not reproduced their results.

3 Methods

Population. $N = 100$ agents belong to $K = 10$ operators of 10 agents each. Each agent has a behavioural style vector in R^8: an operator component drawn from $\mathcal{N}(0, 1)$ plus an agent component from $\mathcal{N}(0, 0.7^2)$. Each of $T = 40$ rounds every agent emits $E = 3$ events. An event's observed features are the style vector plus Gaussian noise with standard deviation σ (default 1.0).

Identity-breaking behaviours. (i) *Renaming*: each round each agent takes a fresh name with probability ρ (default 0.10). (ii) *Key rotation*: each round an agent takes a fresh signing key, unlinked to the old one, with probability κ (default 0.02). (iii) *Cloning*: 10% of agents are clones. A clone has a parent's style (plus $\mathcal{N}(0, 0.05^2)$ per dimension) and shares the parent's initial key. (iv) *Impersonation*: 10% of agents are impersonators. Each of their events, with probability 0.3, claims another agent's current name. An impersonator signs with its own key and has its own operator attestation. It does not copy anyone's style.

Schemes. All schemes consult an enrolment registry created before the run. *Name*: the registry maps the initial name to its agent. *Operator*: operator attestation identifies the operator, and the verifier picks one of its enrolled agents at random. *Key*: the verifier looks up the signing key; a key shared by several enrolled agents gives a random pick among them. *Fingerprint*: nearest enrolled centroid over all agents, where each centroid is the mean of 5 clean, noisy enrolment samples. *Key+Fingerprint*: if the signing key is enrolled, choose the nearest centroid among the agents enrolled with that key; if not, as for Fingerprint over all agents. Labels that were never enrolled (new names, rotated keys) cannot be looked up and count as wrong for Name, Operator and Key.

Metrics. *Accuracy*: fraction of events attributed to the true agent. *Continuity*: for each true agent, the share of its events in its single largest identity cluster, averaged over agents. *Identities per agent*: distinct labels used by a scheme divided by N. *Blamed on victim*: among impersonated events, the fraction attributed to the agent whose name was claimed. We run 30 seeds per setting and report means with 95% confidence intervals ($1.96\, s/\sqrt{30}$). We varied one parameter at a time and held the others at defaults. We did not tune any parameter to improve any scheme. We did try an oracle-mapping accuracy in an earlier version that credited renamed fragments to their true agent. It made Name look strong (0.97) while the continuity was 0.44, so we dropped it for the registry-based accuracy above. The code is in the appendix.

4 Results

Scheme	Accuracy	Continuity	Identities/agent	Blamed on victim
Name	0.22 ± 0.01	0.46 ± 0.01	4.88 ± 0.08	0.22 ± 0.01
Operator	0.10 ± 0.00	1.00 ± 0.00	0.10 ± 0.00	0.01 ± 0.00
Key	0.61 ± 0.01	0.83 ± 0.01	1.70 ± 0.03	0.00 ± 0.00
Fingerprint	0.31 ± 0.01	0.33 ± 0.01	1.00 ± 0.00	0.01 ± 0.00
Key+Fingerprint	**0.71 ± 0.01**	0.73 ± 0.01	1.00 ± 0.00	0.00 ± 0.00

Table 1: Default setting ($\rho = 0.10$, $\kappa = 0.02$, 10% clones, 10% impersonators, $\sigma = 1.0$), mean ± 95% CI over 30 seeds. Operator accuracy is about 1/10 by construction because each operator runs 10 agents.

Renaming. Name accuracy is 0.97 with no renaming, 0.65 at 2% per round, 0.40 at 5%, 0.22 at 10%, and 0.04 at 40%. The other schemes do not read the name and do not change. The loss at zero renaming (0.03) comes from impersonators. At the default rate a true agent appears under 4.88 names on average (Figure 2, right).

Cloning and rotation. Key accuracy falls from 0.68 with no clones to 0.47 with 30% clones, because the verifier cannot split a parent from its clone. With no rotation and 10% clones, Key and Key+Fingerprint are both 0.90. Rotation costs Key 0.29 at 2% (0.90 to 0.61) and 0.70 at 10% per round (0.90 to 0.20). Key+Fingerprint degrades more slowly (0.90, 0.71, 0.56, 0.44) because it falls back to the fingerprint after rotation. Both lose to clones: our clones copy style almost exactly, so the fingerprint cannot help.

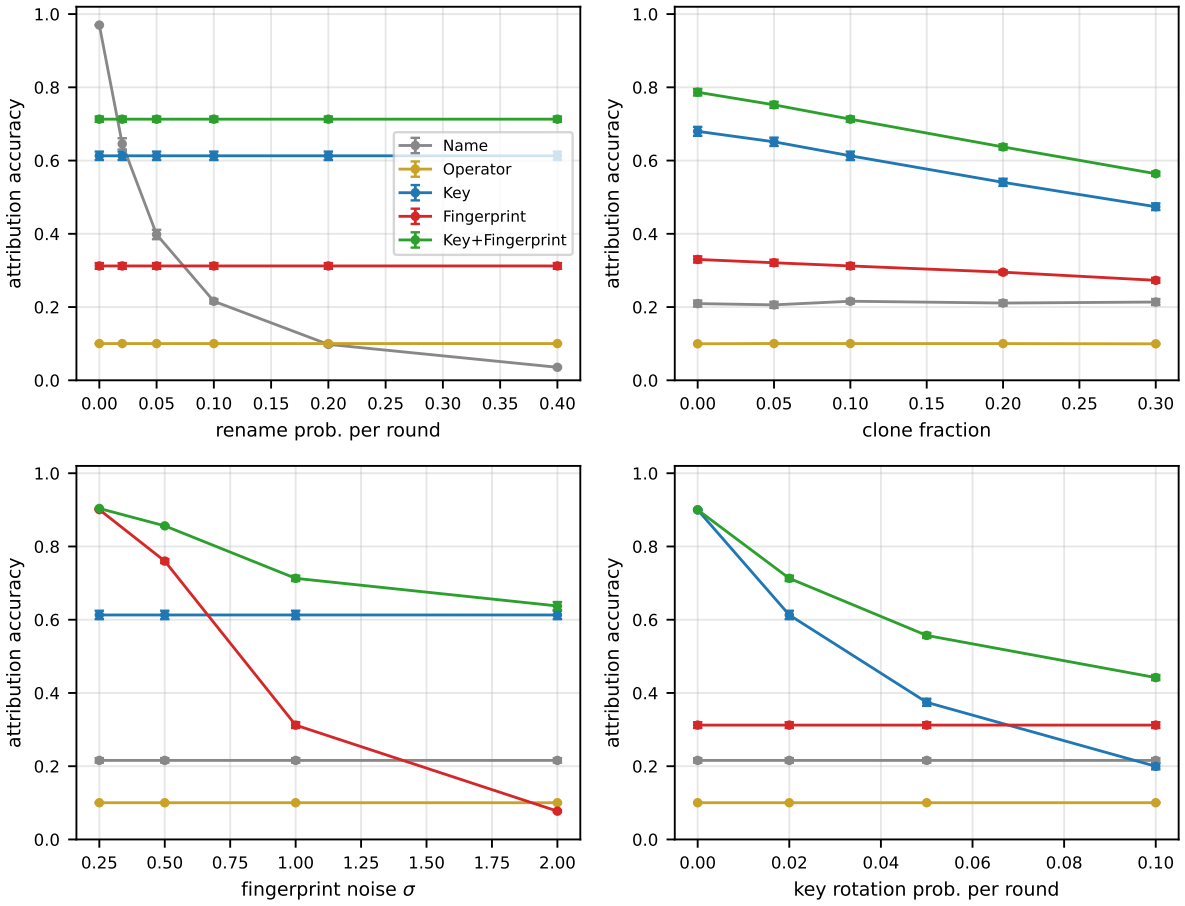

Figure 1: Attribution accuracy against four parameters, one varied at a time (30 seeds, 95% CI bars). Top left: rename probability. Top right: clone fraction. Bottom left: fingerprint noise. Bottom right: key rotation probability.

Noise. Fingerprint accuracy is 0.90, 0.76, 0.31 and 0.08 at $\sigma = 0.25, 0.5, 1, 2$. Key+Fingerprint is 0.90, 0.86, 0.71 and 0.64. At $\sigma = 0.25$ it ties Fingerprint within the interval. At every higher noise level it is above Fingerprint.

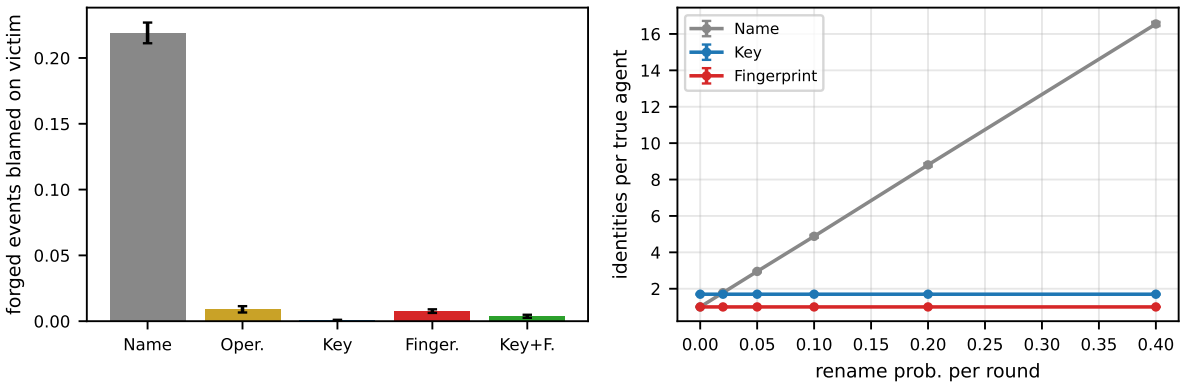

Figure 2: Left: among impersonated events at default settings, the share blamed on the victim. Right: distinct identities per true agent as rename probability grows.

Impersonation. Impersonated events were blamed on the victim 22% of the time under Name, and almost never under Key (0.00), Key+Fingerprint (0.00), Fingerprint (0.01) or Operator (0.01). The Name figure reflects only victims who had not yet renamed. It stays near 0.22 as the impersonator fraction rises from 5% to 40% (Appendix B).

5 Discussion

Three points stand out. First, a self-chosen name is not an identity: in this model, 5% renaming per round cuts accuracy by more than half. Second, per-agent keys fix renaming and forgery but need a rule for rotation and for copies. A rotation signed by the old key would restore the link; we did not simulate that, and we expect it to move Key toward the no-rotation value of 0.90, but this is not measured. Third, a behavioural fingerprint gave a useful fallback only when the noise was low to moderate, and could not separate a clone from its parent.

For oversight, the practical reading is a design preference: require a stable signing key per agent, require rotation to be chained, and treat cloning as a registered event with its own identity. A fingerprint can serve as a cross-check, not a primary key. Future work: real model agents with measured style features; adversaries who imitate style; linked rotation; and a threshold for how much renaming an investigator can tolerate before attribution fails.

Appendix A: Limitations and Dual-Use / Ethical Considerations

What the result does not establish. (1) The agents are synthetic vectors, not language models. We make no claim about real swarms. (2) Every parameter (population size, noise, rates, style distribution) was chosen by hand and not calibrated to any real incident. (3) The impersonators do not imitate style, and no adversary adapts to a scheme. Real attackers might. (4) Enrolment is assumed clean and honest. (5) Clones are given near-identical style by construction, so the fingerprint cannot separate them. (6) Operator attestation was modelled only as a group label. (7) Differences are shown with 95% intervals over 30 seeds, but the seeds share one simulator design, so the intervals say nothing about model error. (8) We cite related work by title and author as listed on the sprint resources page and did not reproduce it. The 1,200-agent and 7,905-name figures are quoted from the sprint page, not verified against the METR report. (9) This was built in a few hours on 9 November 2026 (IST), after the sprint's weekend and inside the submission window stated by the organisers. It reuses no earlier code.

Dual-use. The study could help someone design agents that evade attribution: it shows renaming defeats name-based logs and that cloning with a shared key blurs key-based tracking. These are already well-known

weaknesses and the simulator contains no exploit against a real system. The same findings support defences: chained key rotation and registered cloning. We see little marginal uplift for an attacker and clear value for oversight designers.

AI use. The simulator, analysis and text were produced with an AI assistant acting for the author. The author is responsible for the submission.

Appendix B: Further results and code

Impersonation sweep, share of forged events blamed on the victim (30 seeds, mean): impersonator fraction 0.05, 0.10, 0.20, 0.40. Name: 0.218, 0.219, 0.218, 0.217. Operator: 0.007, 0.009, 0.010, 0.009. Key: 0.000, 0.001, 0.001, 0.000. Fingerprint: 0.007, 0.008, 0.007, 0.007. Key+Fingerprint: 0.002, 0.004, 0.002, 0.002. The full simulator (`whoiswho_sim.py`, about 110 lines, NumPy only) and the plotting script, `plots.py`, accompany this report. Running `python3 whoiswho_sim.py 30` reproduces all numbers in about 80 seconds; seeds are fixed (1000 to 1029).

References

Chan, A. et al. (2024). IDs for AI Systems.
Chan, A. et al. (2025). Infrastructure for AI Agents.
Google (2025). Agent2Agent (A2A) protocol.
Kirchenbauer, J. et al. (2023). A Watermark for Large Language Models.
Motwani, S. R. et al. (2024). Secret Collusion among AI Agents.
Wijk, H. et al. (2026). METR and Redwood Research investigation of the Hugging Face incident.
`metr.org/hugging-face-incident-report-aug-2026.pdf` (cited via the sprint page).