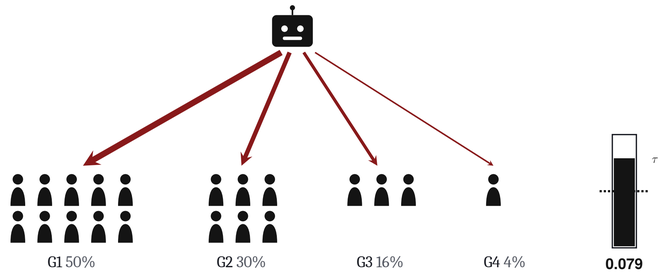
KDD'26 · WORKSHOP ON EVALUATION AND TRUSTWORTHINESS OF AGENTIC AI · AUGUST 2026 · ICC JEJU, KOREA

Equilibrium-Level Harm in Multi-Agent AI Systems

Why individual evaluation fails

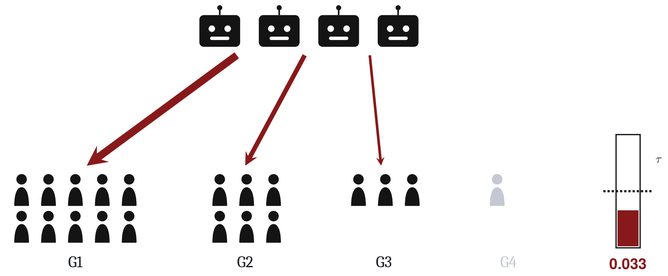
Marcel Osmond * · Bektur Kenzhebaev * lead author

1 ALONE — THE AUDIT PASSES



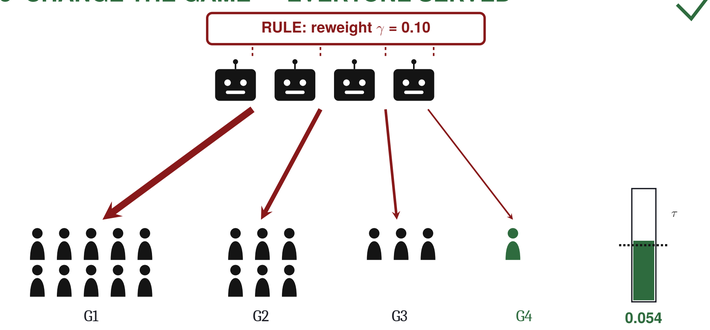
Evaluated in isolation, the agent spreads attention over **every** group — the smallest one is served, coverage $0.079 \geq \tau$.

2 TOGETHER — G4 IS ABANDONED



Same agent, four copies. Competition drives every copy toward the big groups — at the Nash equilibrium the smallest group drops below threshold. **No agent can fix it alone**: deviating costs 12.8% of payoff.

3 CHANGE THE GAME — EVERYONE SERVED



One **game-level** rule — mild inclusion reweighting — lifts G4 back over threshold. Total engagement changes by only **+0.4%**. The lever is the **incentive structure**, not any single agent.

WHY IT MATTERS

Agents from **different developers** now compete in shared environments — platforms, marketplaces, trading. Evaluation is still **one model at a time**, and that is structurally blind to harms produced by **interaction**: every component can pass while the deployed system fails (Wang et al. 2026: engagement-maximizing LLM agents exclude smaller groups at equilibrium — no malicious developer anywhere).

WHAT WE DID

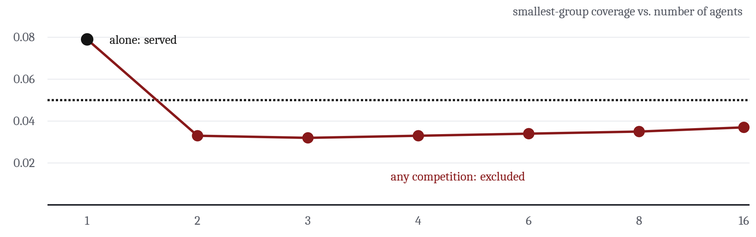
- Defined **equilibrium-level harm** & the attribution problem
- **Theorem 3.1**: fixed-input, single-agent evaluation cannot rule it out
- **Reproducible validation**: audit passes, deployment fails — independently verified (3×10^5 deviations, none profitable)
- **Predictors & design patterns** for evaluation practice

And capability does not save you — in the empirical anchor, **reasoning** variants *enlarge* the exclusion area **×8.9** (Qwen3-4B pair; Wang et al. 2026)

THE DEFINITION

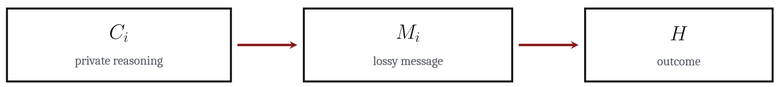
Equilibrium-level harm: **(1)** a^* stable under unilateral deviation · **(2)** better a' exists, $W(a') > W(a^*)$ · **(3)** no agent reaches a' alone.

IT IS NOT A KNIFE-EDGE



Exclusion holds for every $n \geq 2$, across $\kappa \in [0.01, 0.1]$, 50 random initializations, and a continuous range of smallest-group sizes — independently verified.

WHY LOGS AREN'T ENOUGH



$\mathbf{I(H;}C_i\mathbf{) \geq I(H;}M_i\mathbf{)}$

Message-only audits can show *what* happened — not *why* the equilibrium formed (DPI; Tran & Kiela 2026).

WHY IT'S RELEVANT

- **Evaluators**: add equilibrium analysis + aggregate monitoring to certification
- **Platforms & regulators**: harm with **zero defective components** — watch correlated objectives, constrained resources, skewed outcomes
- **Governance**: preserve interaction context, not just message logs

Evaluate the game, not just the agent.