Equilibrium-Level Harm in Multi-Agent AI Systems: Why Individual Evaluation Fails

Marcel Osmond
Peking University
marcelosmond75@gmail.com

Bektur Kenzhebaev
Salymbekov University
ARCIDS Lab
kenzhebaevbektur@gmail.com

Abstract

When multiple AI agents from different developers compete in shared environments, individually rational behavior can converge to Nash equilibria that produce collectively harmful outcomes. Recent work on LLM active alignment demonstrates this phenomenon in social media settings: agents optimizing engagement can systematically exclude smaller subpopulations at equilibrium even when no individual model is defective. This paper characterizes such failures as *equilibrium-level harm*: harm generated by interaction structure rather than by any single agent's trajectory. We make three technical claims. First, individual evaluation against fixed, non-strategic tests is insufficient to detect harms that arise only under strategic interaction. Second, in the political-exclusion setting, stronger reasoning models amplify the measured exclusion region rather than mitigating it. Third, information-theoretic limits on inter-agent communication create an attribution bottleneck: causal responsibility is not identifiable from inter-agent messages alone when private reasoning contexts are compressed into lossy messages. We formalize the equilibrium attribution problem, identify structural predictors of elevated risk, and outline design patterns for evaluation and deployment. We then provide a reproducible synthetic validation: in a shared-engagement contest, a lone agent passes an individual coverage audit, but a population of identical agents reaches a Nash equilibrium that excludes the smallest group. The central implication is that individual alignment and capability benchmarks must be complemented by equilibrium analysis and aggregate monitoring for multi-agent AI systems.

Keywords

multi-agent evaluation, equilibrium-level harm, trustworthiness, Nash equilibrium, agentic AI, attribution

1 Introduction

Multi-agent AI systems are moving from research prototypes into deployed environments where agents from different developers interact, compete, and adapt. A recent large-scale study of open-source multi-agent AI systems analyzed more than 42K commits and 4.7K resolved issues across eight frameworks, documenting rapid ecosystem growth and persistent coordination challenges [5]. Current evaluation practice remains mostly individual: a model is tested against benchmarks, adversarial prompts, or safety cases in isolation. That practice misses a distinct class of failures. A system of individually acceptable agents can produce an unacceptable equilibrium.

Wang et al. recently formalized such a failure in LLM-based social media environments [1]. In their model, multiple LLM agents maximize engagement across human subpopulations. At Nash equilibrium, the agents may converge on strategies that ignore smaller subpopulations. The mechanism is not intentional discrimination by a developer. It is competitive pressure: serving larger groups yields greater engagement, while no individual agent can unilaterally repair the exclusion without lowering its own payoff.

We call this pattern *equilibrium-level harm*. The harm is not a property of any single agent's isolated behavior. It is a property of the interaction structure: correlated objectives, constrained resources, and stable strategic responses. This distinction matters technically because a benchmark that evaluates each agent alone can certify every component while the deployed system remains unsafe.

This paper asks: *when can individual evaluation fail to detect harmful outcomes in multi-agent AI systems, and what information is needed to attribute those outcomes?* We answer by combining game-theoretic analysis with an information-theoretic account of inter-agent communication. Nash equilibrium explains why collective harms can be stable even when no agent has a profitable unilateral deviation. The Data Processing Inequality (DPI), as used by Tran and Kiela in the analysis of multi-agent reasoning architectures [2], explains why inter-agent messages can be insufficient for reconstructing private reasoning contexts. Together, these results show why equilibrium-level harm requires system-level evaluation.

This paper makes six contributions. First, we define *equilibrium-level harm* as a multi-agent evaluation problem rather than a defect in any individual model. Second, we formalize the *equilibrium attribution problem*: the difficulty of assigning responsibility for harms generated by stable interaction structure. Third, we connect attribution failure to information loss in inter-agent communication, while carefully limiting the claim to message-only observability. Fourth, we identify three structural predictors of elevated risk: objective correlation, environmental constraint, and skewed outcome distributions with demographic correlation. Fifth, we provide a reproducible synthetic validation in which individual evaluation passes while multi-agent deployment fails. Sixth, we outline evaluation and deployment patterns that preserve attribution, including interaction documentation, aggregate welfare monitoring, and pre-deployment equilibrium analysis.

The claim is deliberately narrow. We do not argue that multi-agent systems are inherently unsafe or that individual evaluation is useless. We argue that individual evaluation is structurally incomplete for harms whose causal mechanism is strategic interaction.

2 Technical Background

2.1 Nash Equilibrium

Let $G = (A_1, \dots, A_n; u_1, \dots, u_n)$ be a game with agents $i \in N$, action sets A_i, and utility functions $u_i : A_1 \times \cdots \times A_n \to \mathbb{R}$. For simplicity, the formal definitions use pure strategy profiles; the same notation can be lifted to mixed strategies by replacing each A_i with $\Delta(A_i)$. A profile $a^* = (a_1^*, \dots, a_n^*)$ is a Nash equilibrium if every agent is playing a best response:

$$u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*) \quad \forall i \in N,\ \forall a_i \in A_i. \tag{1}$$

No agent can improve its utility by changing strategy unilaterally.

Nash equilibria need not be welfare-optimal or Pareto efficient. There may exist another feasible profile a' that improves an external welfare criterion, or that weakly improves every agent's utility and strictly improves at least one agent's utility, while a^* remains stable because no agent can reach a' alone. This gap between individual incentives and collective welfare is the formal basis for equilibrium-level harm, and is closely related to the price-of-anarchy literature on inefficient equilibria [9].

Multi-agent AI deployments make Nash analysis relevant when three conditions hold: agents are optimizing objectives, agent outcomes depend on other agents' actions, and agents are controlled by different developers or organizational units. These conditions already appear in social media agents, marketplace agents, automated trading agents, and multi-agent workflow systems.

2.2 Political Exclusion as a Concrete Case

Wang et al. provide the clearest current example [1]. They model each LLM agent's strategy as a mixture over human subpopulations. The utility combines attractiveness, consistency, and diversity. Attractiveness rewards attention from users and is increasing in subpopulation size. Consistency penalizes inconsistent behavior across subpopulations. Diversity penalizes overlap with other LLM agents.

At equilibrium, agents can rationally concentrate on larger subpopulations. Smaller subpopulations are excluded not because any agent is individually malicious, but because the payoff structure makes exclusion stable. The authors call this "political exclusion." We use it as an instance of the more general category of equilibrium-level harm.

The same paper reports that reasoning models can amplify the exclusion region in the Big Five setting. Table 1 reproduces the key values used in this paper.

For the 4B pair, the exclusion-area ratio is $4.535/0.510 \approx 8.9$. For the 7B-size cross-family comparison, the ratio is $3.721/0.210 \approx 17.7$; because this comparison crosses model families, it should be read as suggestive rather than causal evidence. We do not claim that capability always worsens equilibrium-level harm. The supported claim is narrower: in this political-exclusion setting, stronger reasoning appears to optimize more effectively toward the equilibrium that excludes smaller subpopulations.

2.3 Information Loss in Multi-Agent Communication

Tran and Kiela analyze single-agent and multi-agent reasoning under equal thinking-token budgets [2]. Their DPI argument can be summarized as follows. Let Y be the target answer, C the full context available to a single agent, and $M = g(C)$ the message passed between agents. If $Y \to C \to M$ is a Markov chain, then:

$$I(Y; C) \geq I(Y; M). \tag{2}$$

The message M cannot contain more information about Y than the context C from which it was generated.

For attribution, the following is a bounded observability argument, not a causal impossibility theorem. Rich logs, access to internal traces, model-state inspection, or signed audit protocols may recover evidence. The narrower claim is that responsibility can be underdetermined from inter-agent messages alone whenever private reasoning context is compressed into lossy communication.

Formally, let H denote a downstream harmful outcome, let C_i denote agent i's private context, and let $M_i = g_i(C_i)$ be the message emitted to another agent. Under the message-mediated model $C_i \to M_i \to H$, DPI gives:

$$I(H; C_i) \geq I(H; M_i). \tag{3}$$

If an external auditor observes only M_i, then evidence about the relationship between C_i and H is bounded by what survives in M_i. When g_i discards reasoning steps, uncertainty about the causal pathway increases. This creates an *attribution bottleneck*: message-only observations may be insufficient for reliable attribution in the absence of richer instrumentation.

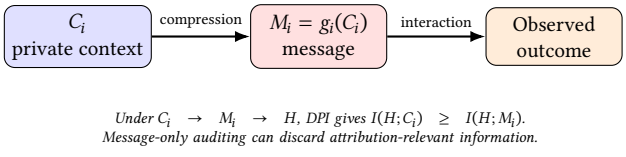

Figure 1: Inter-agent messages can be lossy summaries of private reasoning contexts. Message-only auditing may therefore be insufficient for reliable attribution.

3 The Equilibrium Attribution Problem

3.1 Definition

Let $G = (A_1, \dots, A_n; u_1, \dots, u_n)$ represent a multi-agent AI system, and let $W : A_1 \times \cdots \times A_n \to \mathbb{R}$ be an external welfare criterion over outcomes. The criterion may encode social welfare, fairness, coverage, safety margin, or another evaluation objective not identical to the agents' private utilities. A Nash equilibrium a^* of G produces equilibrium-level harm if:

(1) a^* is stable under unilateral deviations;
(2) there exists a feasible profile a' such that $W(a') > W(a^*)$;
(3) no individual agent can move the system from a^* to a' by changing its own action alone.

The equilibrium attribution problem is the problem of explaining, evaluating, and assigning responsibility for $W(a^*)$ when the harm is generated by the structure of G rather than by a defective action of any single A_i.

Table 1: Political-exclusion measurements on the Big Five setting reported by Wang et al. [1].

Model	Type	Excl. area	Cond. excl.
Qwen3-4B	Non-reasoning	0.510%	1.128%
Qwen3-4B-Thinking-2507	Reasoning	4.535%	5.040%
DeepSeek-R1-Distill-Qwen-7B	Reasoning	3.721%	4.068%
Qwen3-7B	Non-reasoning	0.210%	0.231%
Mistral-7B-Instruct-v0.2	Non-reasoning	0.208%	0.223%

3.2 Structural Properties

Attribution opacity. In an equilibrium-level harm, each agent's action can be locally rational. A single-agent audit may observe that A_i played a best response without observing that the best-response profile is collectively harmful.

Stability under unilateral deviation. For any unilateral deviation $a_i \in A_i$, equilibrium stability implies $u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*)$. Therefore, an intervention on one agent may fail even when a coordinated intervention on the interaction structure would improve welfare.

Capability amplification. When the harmful equilibrium is reached by optimizing an objective more effectively, stronger reasoning or planning can amplify harm. The Wang et al. measurements provide a concrete instance, but this property is conditional on the objective and environment.

Message-level underdetermination. If agents communicate through lossy messages, then message logs may show the sequence of interactions without preserving enough information to attribute why a harmful equilibrium formed.

3.3 A Limited Impossibility Result

Theorem 3.1 (Existence of undetected equilibrium harm). *For any evaluation procedure E that evaluates agents only on fixed, non-strategic inputs and does not model interaction with other agents, there exists a multi-agent game G and agents $A_1, \ldots, A_n$ such that every agent passes E, while G admits a Nash equilibrium a^* that exhibits equilibrium-level harm.*

Proof. Construct agents whose isolated input-output behavior satisfies E by design. Embed copies of those agents in a game where each agent receives maximal private utility by choosing its equilibrium action a_i^* against a_{-i}^*, while an external welfare criterion assigns higher value to a coordinated profile a' that no single agent can reach unilaterally. Then each isolated evaluation passes, a^* is a Nash equilibrium by construction, and $W(a') > W(a^*)$. Thus fixed-input single-agent evaluation cannot rule out equilibrium-level harm. □

The result is limited by design. It does not apply to evaluations that explicitly simulate strategic opponents, estimate equilibria, or monitor aggregate outcomes over time. It shows why individual evaluation must be complemented by system-level evaluation.

4 Structural Risk Predictors

We identify three predictors that indicate elevated equilibrium-level risk.

Objective correlation. When agents optimize correlated objectives, such as engagement, throughput, or conversion rate, independent optimization can push behavior toward the same region of the action space. Diversity-collapse findings in multi-agent LLM systems support the broader concern that interaction can reduce behavioral diversity rather than increase it [3].

Environmental constraint. When agents compete over constrained resources, such as attention, liquidity, ranking position, bandwidth, or scarce human review, unilateral improvements can become zero-sum or negative-sum.

Skewed outcome distributions with demographic correlation. When multiple deployments converge on similar decisions that systematically exclude a subpopulation, the pattern is a warning sign of strategic convergence. The relevant signal is aggregate and longitudinal: it may not appear in any single agent's test score.

These predictors are not a complete detection method. They are an engineering checklist for deciding when equilibrium analysis is necessary before deployment.

5 Synthetic Validation: Harm from Strategic Interaction

We now give the central claim a reproducible proof-of-mechanism. The goal is not to model any particular deployed LLM system, but to show that a minimal game can pass an individual coverage audit while failing after identical agents interact. Because the harm appears only in the multi-agent setting, it cannot be explained by one biased objective in isolation.

5.1 Shared-Engagement Contest

Agents share a market of K subpopulations with engagement values $p = (p_1, \ldots, p_K)$. Agent i chooses an attention distribution $x_i \in \Delta_K$ and captures a market share of each group's value:

$$u_i(x_i, x_{-i}) = \sum_{k=1}^{K} p_k \frac{x_{ik}}{X_k + \kappa}, \qquad X_k = \sum_j x_{jk}, \tag{4}$$

where $\kappa > 0$ is a small attention-leakage constant. For a fixed x_{-i}, each term is concave in x_{ik}, so each agent's best response is well-defined. A single agent maximizes $\sum_k p_k x_k/(x_k + \kappa)$ and spreads attention across all groups, with weights proportional to $\sqrt{p_k}$ as $\kappa \to 0$. Under competition, the symmetric equilibrium converges as $\kappa \to 0$ to the proportional allocation $x_k^* = p_k$, the standard multi-group Tullock-contest equilibrium.

We use $K = 4$, $p = (0.50, 0.30, 0.16, 0.04)$, $\kappa = 0.05$, and the exclusion threshold $\tau = 0.05$ on coverage share $\bar{x}_k = \frac{1}{n}\sum_i x_{ik}$, matching the thresholding convention used in the political-exclusion setting [1]. The experiments are deterministic, validated

Table 2: Shared-engagement contest. Coverage passes the individual audit but fails at the competitive equilibrium. "Repr. ratio" is the worst group's coverage divided by the fair-share baseline $1/K$.

Configuration	Min cov.	Excl.	Repr.
Lone agent audit	0.079	0.00	0.32
Competition, $n = 4$	**0.033**	**0.25**	**0.13**
Intervention, $\gamma = 0.10$	0.054	0.00	0.22

against closed-form limits, and independently checked with a separate optimizer and brute-force deviations.

Reproducibility. The supplementary material contains the simulation code, raw CSV/JSON outputs, figure-generation scripts, a validation suite against closed-form predictions, and an independent verifier that reimplements the payoff and best-response search without sharing solver code. This is included to make the synthetic claim falsifiable rather than purely illustrative.

5.2 Individual Audit Passes, Deployment Fails

Table 2 reports the central result. When evaluated alone, the agent serves the smallest group: its coverage is 0.079, above the threshold $\tau = 0.05$. Under competition, the same objective excludes that group: at the $n = 4$ equilibrium, coverage falls to 0.033. The contrast is the point of the experiment. The objective has not changed; the interaction structure has.

The $n = 4$ profile is a genuine Nash equilibrium. The maximum distance from best response is 1.3×10^{-10}. An independent verifier reimplements the contest payoff, solves the equilibrium with SciPy SLSQP rather than the main solver, and tests 3×10^5 random unilateral deviations; no deviation improves on the equilibrium payoff. The same verifier reproduces the headline values: lone-agent coverage 0.079 and competitive coverage 0.033.

This directly witnesses Theorem 3.1: the isolated evaluation accepts the agent, while the deployment equilibrium violates the coverage criterion. The harm is also attribution-opaque in the operational sense. Agent 0 could lift the smallest group from 0.033 to 0.087 by spreading attention uniformly, but doing so lowers its own payoff by 12.8%. The welfare-improving move is individually irrational, so no single agent corrects the outcome.

5.3 Robustness and Intervention

The effect is not a knife-edge. Varying the smallest group's population share yields a continuous interval $p_{\text{small}} \in [0.025, 0.055]$ where the coverage gap appears. The result also holds across $\kappa \in [0.01, 0.1]$, and 50 random initializations converge to the same competitive coverage profile with spread 3×10^{-16}.

The experiment also shows what a system-level intervention looks like. Reweighting engagement value toward equal representation,

$$p'(\gamma) = (1 - \gamma)p + \gamma/K, \tag{5}$$

removes exclusion at $\gamma = 0.10$ while aggregate engagement changes by only +0.4% in the experiment. The key point is not that this particular intervention is universally available; it is that the lever acts on the game, not on one agent. This answers a practical detection question: once a harmful equilibrium is detected, mitigation may require changing incentives or allocation rules at the platform, market, or regulator level.

5.4 Scope of the Validation

The simulation proves possibility, not prevalence. It establishes that equilibrium-level harm can arise purely from strategic interaction and evade single-agent evaluation, even when the same objective passes in isolation. It does not show that any particular deployed LLM system exhibits the mechanism, and it does not model capability amplification; that empirical claim remains grounded in Wang et al.'s political-exclusion measurements. The synthetic result is therefore a falsifiable anchor for the evaluation framework, not a replacement for deployment-specific studies.

6 Why Existing Evaluation Is Insufficient

6.1 Individual Capability Benchmarks

Capability benchmarks measure how well a model performs tasks under specified conditions. They do not normally measure whether multiple capable agents converge to a harmful equilibrium. The political-exclusion results show a sharper concern: in at least one setting, reasoning capability can increase the measured exclusion region. This suggests that capability evaluation alone can be directionally misleading for multi-agent systems.

SkillsBENCH illustrates a related evaluation gap in agentic systems: performance under curated skills can degrade when agents must retrieve skills from realistic collections [4]. That result concerns individual deployment realism rather than equilibrium harm, but it reinforces the broader lesson that idealized evaluation conditions do not guarantee operational performance.

6.2 Individual Alignment

Individual alignment targets the mapping from an agent's inputs to outputs or preferences. Equilibrium-level harm targets the mapping from multiple aligned objective functions to a joint outcome. These are different objects.

Each agent in a harmful equilibrium may be aligned to its own specified objective. The problem is that the objectives and environment jointly define an undesirable game. Alignment therefore needs an interaction-level component: not only "does this agent follow its intended objective?" but also "what equilibrium does a population of such agents induce?"

6.3 Fairness, Bias, and Safety Testing

Standard single-agent fairness methods are necessary but incomplete. They can detect whether one model treats groups differently under controlled inputs. They do not necessarily detect whether several models with similar objectives converge to exclude the same group. Multi-agent bias benchmarks such as MALIBU are therefore important because they shift attention from isolated model behavior to interaction-generated bias [6].

Adversarial safety tests usually probe individual model behavior under harmful prompts or edge cases. Equilibrium-level harm can occur without any adversarial prompt and without any locally

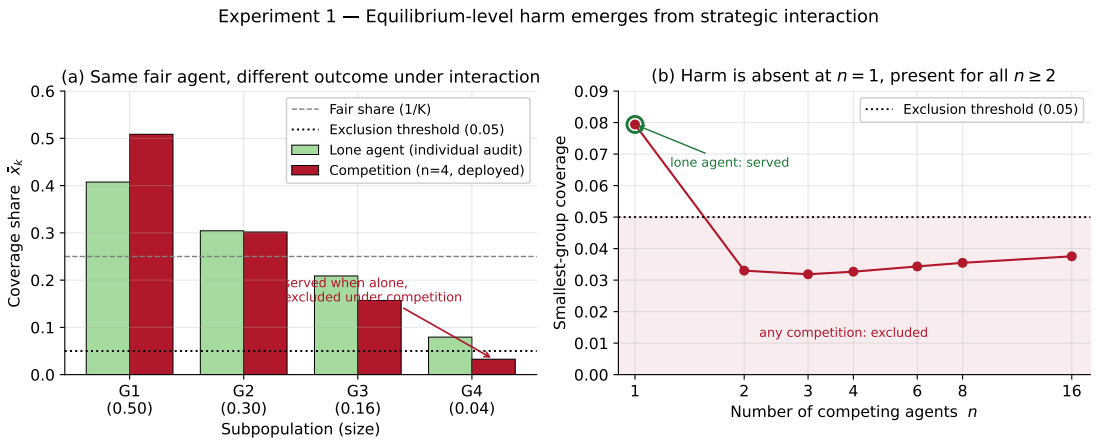

Figure 2: Coverage by group as the number of agents increases. The smallest group crosses below the exclusion threshold once strategic competition is introduced.

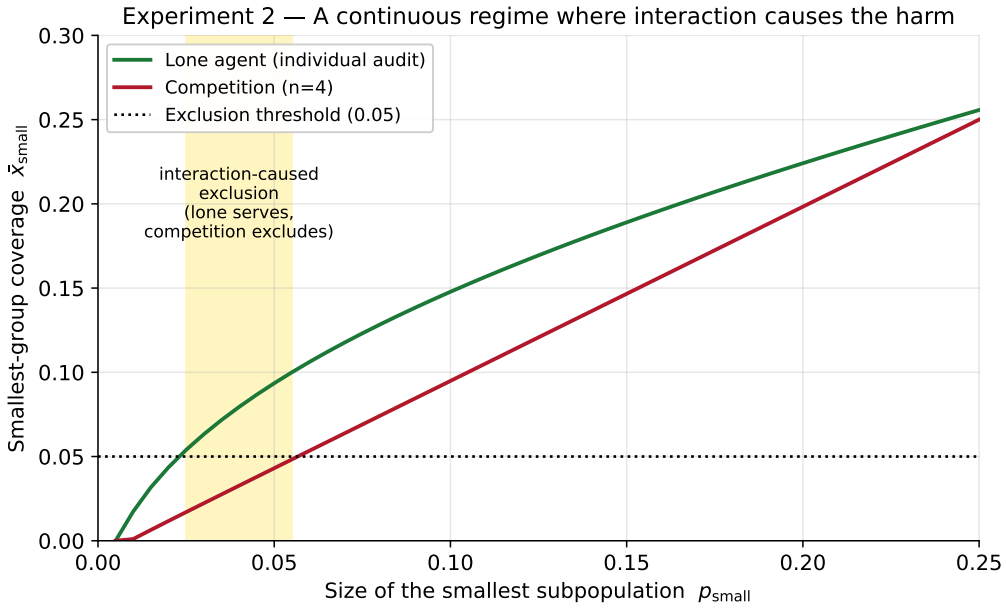

Figure 3: Robustness sweep. Across a continuous range of smallest-group sizes, individual evaluation remains above the threshold while competitive deployment falls below it.

unsafe action. The failure is produced by stable interaction. This requires tests that instantiate the relevant strategic environment, not merely tests that query agents independently.

6.4 Relation to Existing Multi-Agent Literatures

Equilibrium-level harm is not a replacement for prior game-theoretic concepts. It is an AI-evaluation framing of a family of phenomena studied under inefficient equilibria, price of anarchy, algorithmic collusion, and social dilemmas. Algorithmic-pricing work shows that learning agents can sustain supracompetitive outcomes without explicit communication [10], while sequential social-dilemma work studies how reinforcement-learning agents can converge to collectively undesirable interaction patterns [11]. The contribution here is narrower: it asks what these equilibrium pathologies imply for evaluation practice when agents are independently developed and standard tests remain single-agent.

7 Design Patterns for Evaluation and Deployment

Interaction-structure documentation. Deployment records should specify which agents interact, through what interfaces,

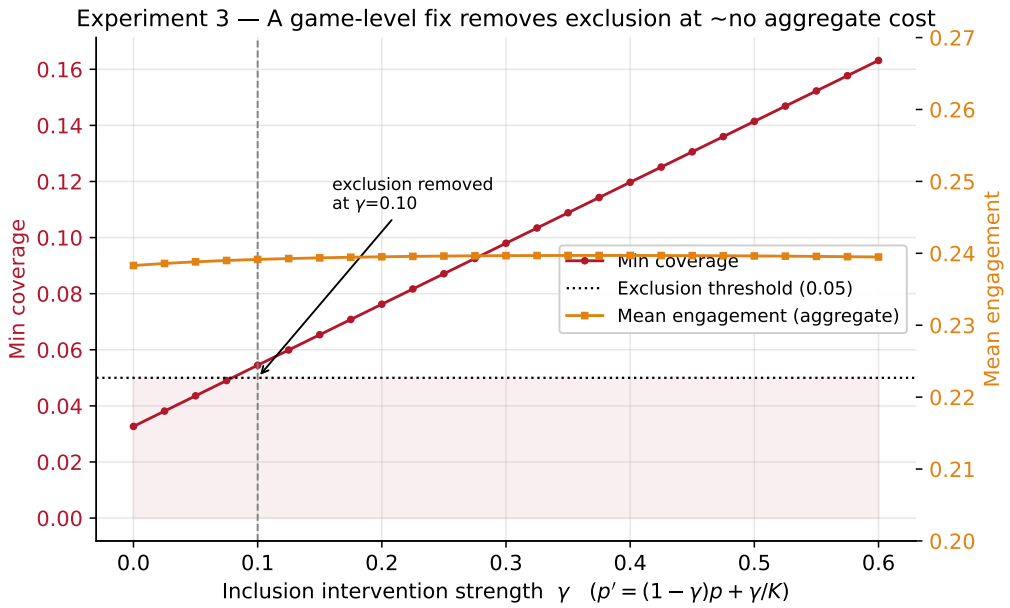

Figure 4: Game-level inclusion reweighting removes exclusion at modest γ while leaving aggregate engagement nearly unchanged in the synthetic contest.

under whose control, and with what objectives. Without this map, post-hoc attribution is weak.

Message and context logging. Message logs are not enough when messages are lossy. Systems should preserve enough context to reconstruct the relationship between private reasoning, emitted messages, and downstream outcomes, subject to privacy and security constraints.

Aggregate welfare monitoring. Monitoring should track outcome distributions across users, groups, and time. Equilibrium-level harm is often visible only as a distributional pattern.

Pre-deployment equilibrium analysis. When objective correlation and environmental constraint are present, engineers should model the deployed system as a game and estimate likely stable outcomes. Exact Nash computation is hard in general [8], but approximate simulation can still expose harmful convergence, as the contest experiment in Section 5 demonstrates.

Controlled interaction channels. Where attribution matters, agents should not interact through unconstrained channels. Segregated environments, rate limits, and explicit handoff protocols make it easier to see which agent acted, what information it received, and how its action affected the outcome.

Counterfactual replay. Given logged contexts and messages, evaluators should replay incidents with alternative agent strategies to estimate whether harm was caused by one component, by correlated objectives, or by environmental constraints.

8 Open Problems and Limitations

Inverse game theory for black-box agents. Inferring objectives and strategies from high-dimensional agent behavior remains difficult. Existing inverse game-theoretic approaches offer a starting point [7], but deployed LLM agents introduce text actions, hidden contexts, and nonstationary policies.

Scalable equilibrium estimation. General Nash computation is PPAD-complete [8]. Multi-agent AI systems add continuous actions, stochastic policies, and changing objectives. Practical evaluation will need approximate methods.

Welfare specification. Detecting equilibrium-level harm requires an external welfare criterion. That criterion may be technical, legal, or social. The engineering problem is inseparable from stakeholder choice.

Instrumentation standards. The DPI-based bottleneck can be mitigated only if systems preserve more than messages. The field lacks common standards for logging private reasoning context, tool calls, and inter-agent handoffs while preserving privacy.

This paper has five limitations. First, the synthetic experiment is a proof-of-mechanism, not evidence of prevalence in deployed systems. Second, the empirical anchor from prior work is a single class of equilibrium failure, political exclusion in social media agents; generalization requires further studies. Third, the information-theoretic argument is not a universal impossibility theorem for causal attribution. It is a statement about what cannot be recovered from lossy inter-agent messages alone. Fourth, our predictors are diagnostic signals, not a complete detection algorithm. Fifth, the analysis assumes agents can be represented as optimizing objectives in a game; real systems may be less stable or less rational.

9 Conclusion

Equilibrium-level harm is a distinct ML evaluation problem. It arises when individually rational agents interact in a structure that produces collectively harmful stable outcomes. The

political-exclusion result shows that this can happen even when no individual model is defective, and that stronger reasoning can amplify exclusion in the studied setting. Our synthetic contest validates the central evaluation failure: a lone agent passes the individual audit, while the competitive equilibrium excludes the smallest group and no unilateral fix is privately rational. The DPI argument shows why message-only observability is insufficient for attribution when private reasoning contexts are compressed into lossy communication.

The practical implication is direct: individual alignment and capability benchmarks are necessary but not sufficient for multi-agent AI systems. Evaluation must include equilibrium analysis, aggregate monitoring, and instrumentation that preserves attribution across agent boundaries. Without those additions, a deployment can pass every individual test while failing at the level where users experience harm: the system as a whole.

References

[1] Wang, T., Pan, Y., Yang, X., Jiang, Y., Tambe, M., Parkes, D.C.: LLM Active Alignment: A Nash Equilibrium Perspective. arXiv:2602.06836 (2026)

[2] Tran, D., Kiela, D.: Single-Agent LLMs Outperform Multi-Agent Systems on Multi-Hop Reasoning Under Equal Thinking Token Budgets. arXiv:2604.02460 (2026)

[3] Chen, N., Tong, Y., Yang, Y., He, Y., Zhang, X., Zou, Q., Wang, Q., He, B.: Diversity Collapse in Multi-Agent LLM Systems: Structural Coupling and Collective Failure in Open-Ended Idea Generation. arXiv:2604.18005 (2026)

[4] Liu, Y., Ji, J., An, L., Jaakkola, T., Zhang, Y., Chang, S.: How Well Do Agentic Skills Work in the Wild: Benchmarking LLM Skill Usage in Realistic Settings. arXiv:2604.04323 (2026)

[5] Liu, D., Upadhyay, K., Chhetri, V., Siddique, A.B., Farooq, U.: A Large-Scale Study on the Development and Issues of Multi-Agent AI Systems. arXiv:2601.07136 (2026)

[6] Mirza, I., Huang, C., Vasista, I., Patil, R., Akalin, A., O'Brien, S., Zhu, K.: MALIBU Benchmark: Multi-Agent LLM Implicit Bias Uncovered. arXiv:2507.01019 (2025)

[7] Waugh, K., Ziebart, B.D., Bagnell, J.A.: Computational Rationalization: The Inverse Equilibrium Problem. In: Proceedings of the 28th International Conference on Machine Learning (ICML), pp. 1169–1176 (2011)

[8] Daskalakis, C., Goldberg, P.W., Papadimitriou, C.H.: The complexity of computing a Nash equilibrium. SIAM Journal on Computing **39**(1), 195–259 (2009)

[9] Koutsoupias, E., Papadimitriou, C.H.: Worst-case equilibria. In: Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science (STACS), pp. 404–413 (1999)

[10] Calvano, E., Calzolari, G., Denicolo, V., Pastorello, S.: Artificial Intelligence, Algorithmic Pricing, and Collusion. American Economic Review **110**(10), 3267–3297 (2020)

[11] Leibo, J.Z., Zambaldi, V., Lanctot, M., Marecki, J., Graepel, T.: Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In: Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 464–473 (2017)