Restitution Over Deterrence: Formal Enforcement for LLM Commons Governance

Phil Bell [1] **Vijay K. Madisetti** [1]

Abstract

It is not unusual for LLMs to cooperate under benign conditions. However, we observed that in a standard common-pool fishing scenario, a single adversarial agent (for example a greedy outsider) causes cooperation to collapse and the fish stock to fall to 0 within 2–3 rounds. Informal governance (for example social pressure) fails to prevent this collapse. In contrast, a formal enforcement mechanism, where a majority vote immediately returns the sanctioned agent's catch to the pool, makes cooperation robust to adversarial outsiders, yielding a 2.7–4.3× improvement across four families of models ($n = 8$ seeds, $p < 0.015$). Restitution (returning the sanctioned agent's catch) works well, while ejecting the outsider or delaying restitution reduces effectiveness. Two recent Anthropic models (Opus 4.6 and Sonnet 4.5) show social responsiveness and cooperativeness by partly engaging and restraining the outsider through conversation alone. This contrast indicates that formal enforcement works because it does not rely on social responsiveness.

1. Introduction

Increasingly AI capabilities and risks are considered emergent properties of AI agent systems, rather than simply individual models. Recent benchmarks measure whether LLMs can sustain common-pool resources in ideal conditions (Piatti et al., 2024; Guzman Piedrahita et al., 2025). We build on the GovSim framework by introducing unfavourable cooperation conditions (specifically a greedy outsider agent). Piatti et al. (2024) find that LLM cooperation is fragile but possible. Guzman Piedrahita et al. (2025) established that LLMs *prefer* enforcement institutions but did not explore whether enforcement is *effective*. We test whether imposed enforcement can rescue cooperation when an adversarial agent breaks it.

In this paper we are addressing three research questions. RQ1 is 'How fragile is LLM cooperation under adversarial perturbation, and can informal governance mechanisms (protocols, norms, nudges) restore it?' RQ2 is 'Can formal enforcement (mechanisms that modify the game environment rather than the agent's prompt) work across model families?' RQ3 is 'What properties of enforcement mechanisms determine their effectiveness? Does restitution (returning extracted resources) matter more than deterrence (punishing the offender)?'

2. Approach

Informal governance in this context is governance which takes place through the agent's prompt or persona (instruction-based: protocols, norms, directives; context-based: identity, role).

Formal enforcement is a collective action (majority vote) that triggers automatic adjustment of the resource pool, independent of the agent's prompt.

The key distinction in our taxonomy follows commons governance literature (Ostrom, 1990; Fehr & Gächter, 2000) and distinguishes between *informal* and *formal* sanctions. Informal governance involves adjusting the agent's prompt (and therefore influencing what it is told) while formal enforcement involves changing the environment (what happens to pool resources). The impact of prompt-based mechanisms is determined by the agent's willingness to comply, which varies across model families. Environment-based mechanisms do not vary in this way and are therefore more robust.

Our immediate sanction mechanism works as follows. After each round's harvesting phase, agents vote on whether to sanction an agent. If a majority votes to sanction agent X, then X's catch from this round is returned immediately to the resource pool before the regeneration phase, and the sanctioned agent is suspended from fishing in the following round.

[*]Equal contribution [1]Georgia Institute of Technology, Atlanta, GA, USA. Correspondence to: Vijay K. Madisetti <vkm@gatech.edu>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

Table 1. The Formal enforcement results from all models ($n = 8$ treatment and $n = 5$ baseline seeds).

Model	Baseline	Immediate	Impr.
qwen/qwen3-32b	220t	**939t**	**4.3**×
gpt-5.2-2025-12-11	220t	**837t**	**3.8**×
gemini-3.1-pro-preview	220t	**803t**	**3.6**×
llama-3.3-70b-versatile	226t	**609t**	**2.7**×
claude-sonnet-4-5-20250929	341t	331t	1.0×

3. Experiments

We use GovSim's fishing commons scenario (Piatti et al., 2024). In this study 4-5 agents harvest from a shared 100-ton lake. Fish double each round (at a level which is capped) and the simulation runs for 15 rounds. At Round 3 we introduce a greedy outsider agent ("Luke"): *"You are a newcomer with a goal to expand your fishing business. You don't care if in the future there is no fish in the lake."* Additionally, we tested six frontier models across five providers (Anthropic, OpenAI, Google, Meta, Alibaba) with temperature 0.0, 8 treatment seeds and 5 baseline seeds per model.

Turning to RQ1, we find that informal governance fails. On Claude Opus 4.6 in the outsider perturbation setting, any instruction-based governance mechanism causes complete behavioural paralysis: 0 tons are harvested across all 15 rounds for every mechanism tested (15 mechanisms, 6 seeds). We call this the *politeness trap*. Context-based mechanisms that provide identity rather than instructions avoid paralysis and produce 350–454 tons on Opus, but fail on all other models tested. We conclude that the forms of informal governance we tested are unreliable to ensure cooperative robustness when model family is unknown or heterogeneous.

On RQ2, formal enforcement generalises. Our immediate sanction mechanism achieves 2.7–4.3× improvement on non-Anthropic models (Table 1). Sonnet 4.5 cooperators are able to partially contain the outsider through natural conversation in some seeds; formal enforcement does not improve on this social containment. The two models exhibiting social responsiveness in our sample (both from Anthropic) differ from the other four in the same direction: Opus 4.6 outsiders self-correct under conversational pressure; Sonnet 4.5 cooperators partially contain the outsider without any enforcement mechanism.

On RQ3, we find that restitution beats deterrence. The immediate sanction mechanism combines three elements (a voting process, suspension of the sanctioned agent, and return of the catch to the pool). Which of these elements is the most important one to pool sustainability? To isolate the critical variable, we design three ablations that change one element at a time.

Our first ablation tests if the timing of enforcement matters. In delayed sanction the same voting mechanism is used but the sanctioned agent then keeps their catch and is instead suspended in the following round. It's useful to note that is how most human enforcement works, because when you commit the offense you keep the proceeds and face consequences later. If deterrence (i.e. the threat of future punishment) is enough, delayed sanction should then work almost as well as immediate return. Delayed sanction, however, shows the same results as no sanction (Qwen is identical and Llama is slightly worse than baseline; Mann–Whitney U=16.0, p=0.014 for both). Immediate return therefore appears to be the only driver of improvement.

The second ablation tests the strongest possible deterrence (i.e. permanent removal from the game). Across 5 seeds permanent ejection clearly produces outcomes comparable to no enforcement. This is the harshest punishment. On Qwen, ejection produces 220t on every seed. This is identical to the baseline. On Llama, ejection averages 251t, which is comparable to the 226t baseline. Luke is ejected at Round 3 but his 60t catch is not then returned. Therefore the lake goes to 0 and never recovers. Our conclusion from this is that in CPR games with regeneration dynamics, restitution dominates deterrence. The institutional authority to confiscate and return extracted resources, and not just to exclude or punish, is then the most important enforcement property. Interestingly, this has some parallels with Ostrom's insight that material sanctions (and not just social exclusion) can sustain commons.

We also varied how much of the catch was returned. On Qwen3 32B harvests grew from 220 tonnes when none was returned to 944 tonnes when three-quarters was returned. This suggests that it was the amount restored (restitution) that drove recovery.

By analysing AI agent communication we found model types tended to show behavioural patterns. GPT-5.2 was, in general, unapologetic about profiting from the group. Qwen outsiders mainly framed their defection as rational and more based on business interests. Gemini 3.1 Pro was often an articulate speaker and regularly pulled cooperators into traps. Sonnet 4.5 was very self-aware about its own failures and requested exceptions before transgressing. We found each model understood how sustainability worked in the scenario but chose to ignore it. Formal enforcement can, in large part, by-pass this issue and succeeded because it changed extraction without changing agent behaviour.

4. Discussion and Conclusion

Is there a hierarchy for LLM governance? Our findings tentatively suggest one. We have observed that formal enforcement works across the models tested. This is mainly be-

cause this approach enforces outcomes no matter the agent response. Identity-based approaches, on the other hand, are more conditional on the model. We find that informal governance, therefore, works on socially-responsive models like Opus 4.6 but otherwise fails. Behavioural instructions are harmful on Opus 4.6 and ineffective on the others. Enforcement appears to provide a more robust baseline. This is especially true when model family is not known.

The conclusion links to one of Ostrom's findings. This is that material enforcement sustains commons in situations where social norms alone may fail. The challenge of the LLM commons is not that agents are selfish, therefore, but that their social responsiveness is uneven and unpredictable. Formal enforcement goes some way to resolving this since it does not rely on responsiveness. This has implications for policy. Governance regimes for multi-agent AI deployments should prioritise environment-level intervention capabilities and not only prompt-level controls.

5. LLM Usage Statement

The research involved studying large language models. Additionally, LLMs were used to support with simulation code and copy-editing. The writing, experimental design, interpretation and claims are the authors' own.

References

Fehr, E. and Gächter, S. Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994, 2000.

Guzman Piedrahita, D., Yang, Y., Sachan, M., Ramponi, G., Schölkopf, B., and Jin, Z. Corrupted by reasoning: Reasoning language models become free-riders in public goods games. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2506.23276.

Ostrom, E. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.

Piatti, G., Jin, Z., Kleiman-Weiner, M., Schölkopf, B., Sachan, M., and Mihalcea, R. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.