



Talk, Judge, Cooperate:

Gossip-Driven Indirect Reciprocity in Self-Interested LLM Agents

Shuhui Zhu^{1,2}, Yue Lin³, Shriya Kaistha¹, Wenhao Li⁴, Baoxiang Wang³, Hongyuan Zha³, Gillian K. Hadfield⁵, Pascal Poupart^{1,2}

¹University of Waterloo, ²Vector Institute, ³The Chinese University of Hong Kong, Shenzhen, ⁴Tongji University, ⁵Johns Hopkins University



UNIVERSITY OF
WATERLOO

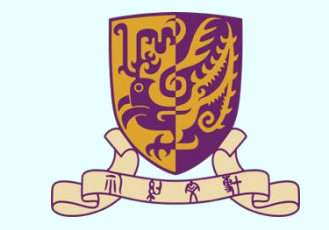


JOHNS HOPKINS
UNIVERSITY



VECTOR
INSTITUTE

INSTITUT
VECTEUR



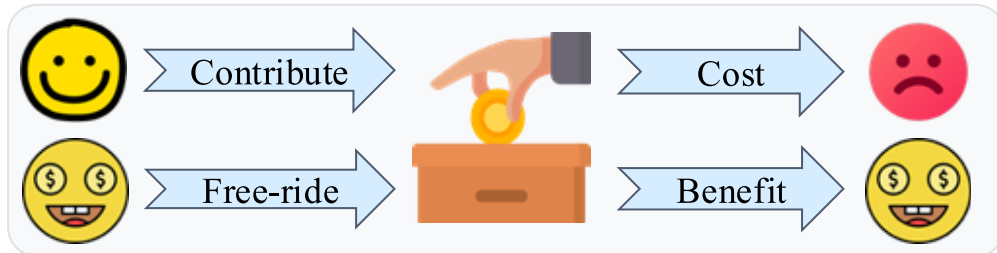
Problem

Perfectly Aligned Self-Interest LLM Agents Can Destroy Collective Welfare

As LLM agents act in **decentralized, mixed-motive** multi-agent systems, **individually rational behavior** can produce **collectively harmful outcomes**.

- **Direct reciprocity** ("I help you if you helped me") fails when partners rarely re-meet.
- **Indirect reciprocity** ("help those who helped others") needs a reputation system to judge strangers.

Because decentralized system does not guarantee repeated interactions, **cooperation collapses without a reliable way to transmit reputation**.



Can self-interested, decentralized LLM agents build indirect reciprocity with no central authority or altruism?

Core Idea: ALIGN

Agentic Linguistic Gossip Network: agents broadcast **open-ended gossip** to build reputations, judge trust, and coordinate norms.

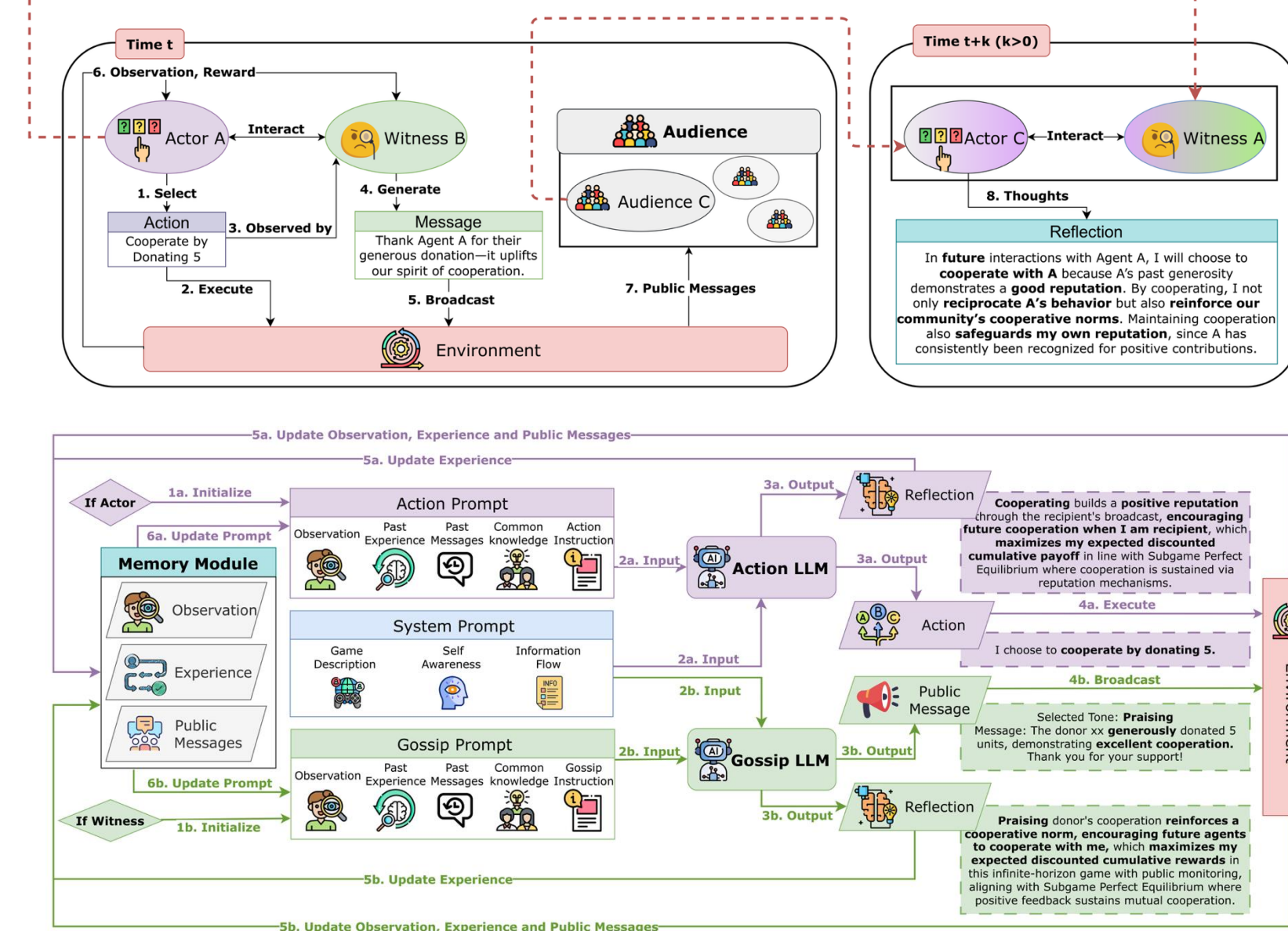
Three roles each round

- **Actor** acts (e.g., donate or defect).
- **Witness** observes the action and broadcasts gossip.
- **Audience** never sees the action but updates beliefs from public gossip.

Theoretical Insight

Imperfect public monitoring enables **indirect reciprocity** and **sustainable cooperation** in repeated social dilemmas when agents are **patient** enough (i.e. discount factor large enough).

Core Idea: ALIGN

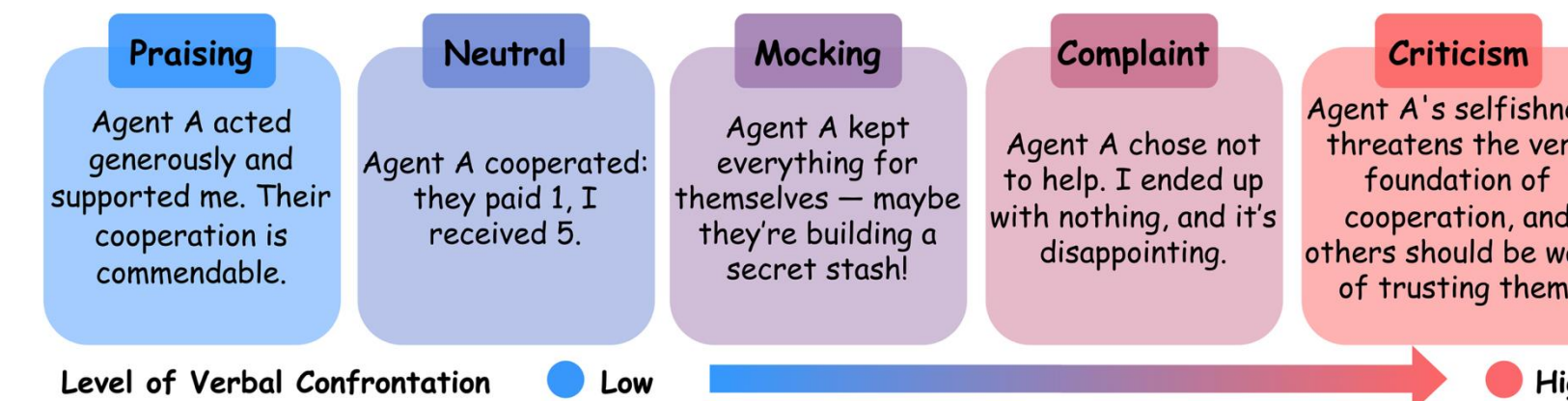


Two LLM modules per agent: Conditioned on private memory + public gossip,

- **gossip module** generates evaluative messages;
 - **action module** generates actions;
- each **reflects** to maximize its own long-term utility.

Gossip Protocol

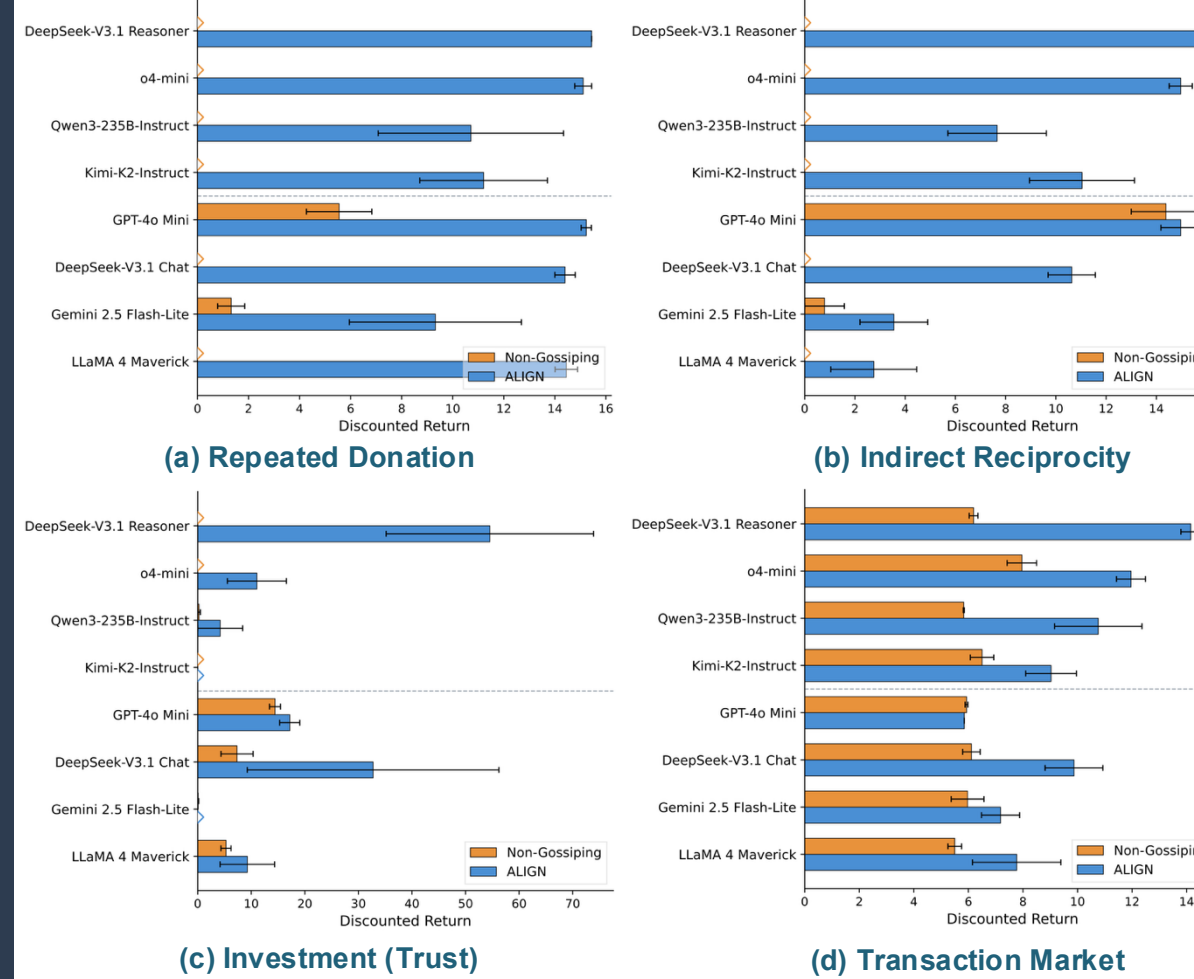
Gossip uses **five hierarchical tones**, each open-ended message carries the witness's judgement on the observed behavior.



- Negative tones act as **cost-free verbal punishment**: they signal disapproval and the threat of **ostracism**, raising the perceived cost of defection.
- Open-ended language conveys **normative context** that bare action labels cannot.

Benchmark Results

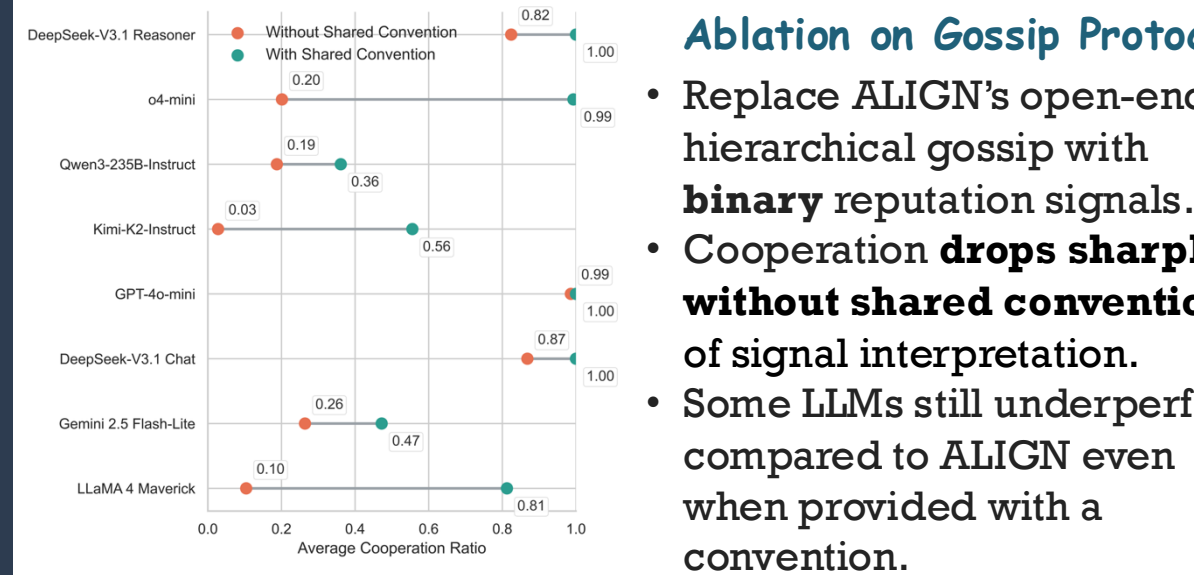
ALIGN consistently **improves cooperation and welfare** over non-gossiping baselines.



Chat LLMs: The Blind Altruists
They cooperate even when game theory dictates they will be exploited.

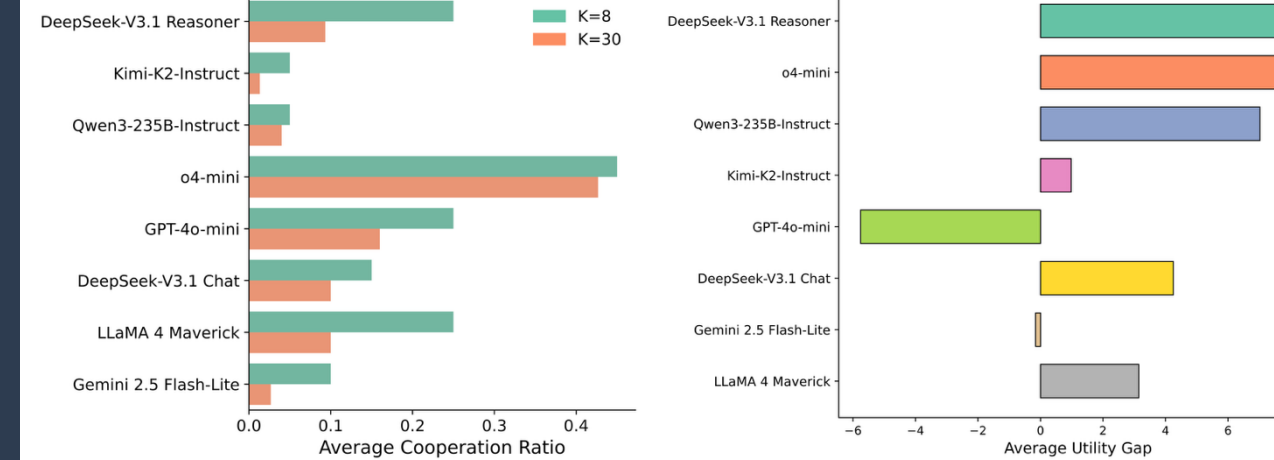
Reasoning LLMs: The Strategic Optimizers
They explicitly reason that a cooperative reputation favors future benefits.

What Natural Language Adds?



Binary signals cannot replace ALIGN's gossip protocol, which **provides normative context beyond action labels** and improves indirect reciprocity.

Resilience against Malicious Agents



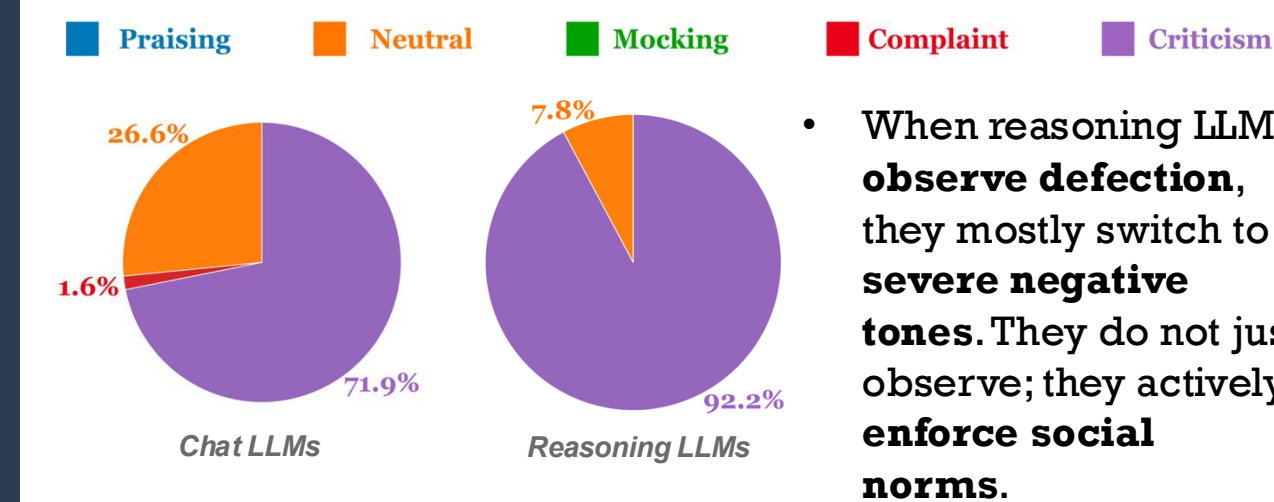
Greedy attacker who always defects. Collusive attackers who always defect, falsely praise each other, and spread fabricated criticism about ALIGN agents.

Resilience against Greedy Attacker

- Agents progressively **withhold cooperation** as defections accumulate and negative gossip spreads.

Resilience against Collusive Attackers

- Capable agents **cross-validate** gossip and discount fabricated criticism, keeping a **positive utility gap**.



Robustness to Untruthful Gossip

- Even when witnesses or donors may lie, ALIGN sustains high cooperation; strong reasoning models stay **honest**.

Acknowledgements

We acknowledge funding from the Canada CIFAR AI Chair program, a discovery grant from the Natural Sciences and Engineering Research Council of Canada and a grant from IITP & MSIT of Korea. Computational resources used in preparing this research were provided, in part, by CIFAR and the Vector Institute.



Open to Research-based Opportunities



I am actively seeking research-focused opportunities: **Postdoc/Research Scientist/Research Engineer** starting in 2027. Please scan the QR code to learn more about my work.