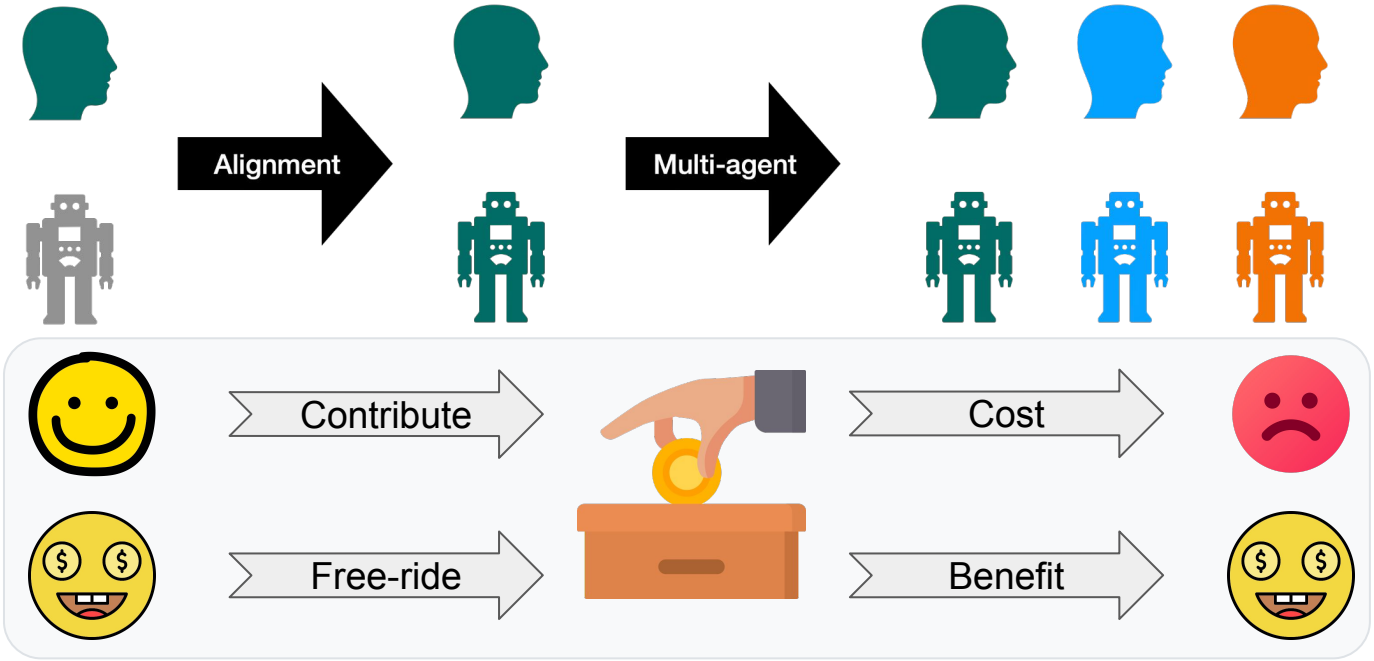
Talk, Judge, Cooperate:
Gossip-Driven Indirect Reciprocity in Self-Interested LLM Agents

Shuhui Zhu, Yue Lin, Shriya Kaistha, Wenhao Li, Baoxiang Wang, Hongyuan Zha, Gillian K Hadfield, Pascal Poupart

Perfectly Aligned Self-Interest LLM Agents Can Destroy Collective Welfare



➔ **Each agent may rationally optimize its own objective.**

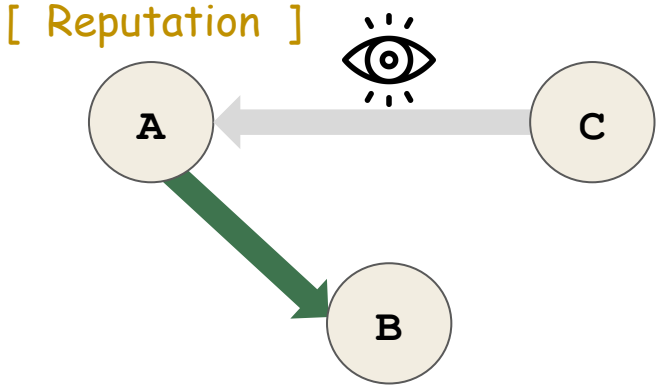
➔ **But mixed incentives can create global failure modes.**

➔ **We need mechanisms that make cooperation strategically sensible.**

Key Insight: When LLM agents deploy in decentralized, mixed-motive ecosystems, individually rational behavior can produce collectively harmful outcomes.

Direct Repayment is Not Reliable in Large Decentralized MAS

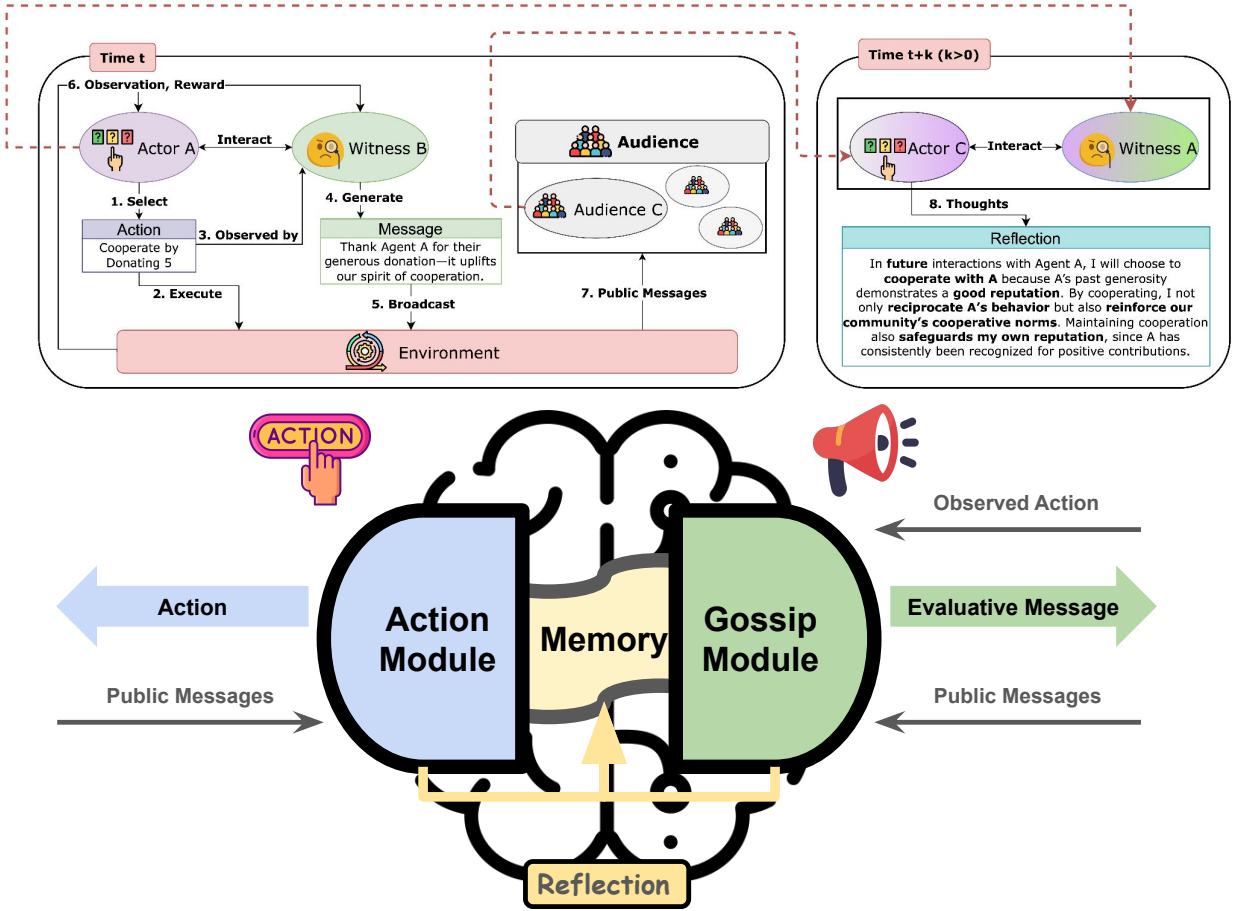
Direct Reciprocity: I help you because you helped me. (Requires repeated encounters).



Indirect Reciprocity: I help you because you helped others. (Requires a reputation mechanism).

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

ALIGN: Public Gossip as a Reputation Mechanism



Action module

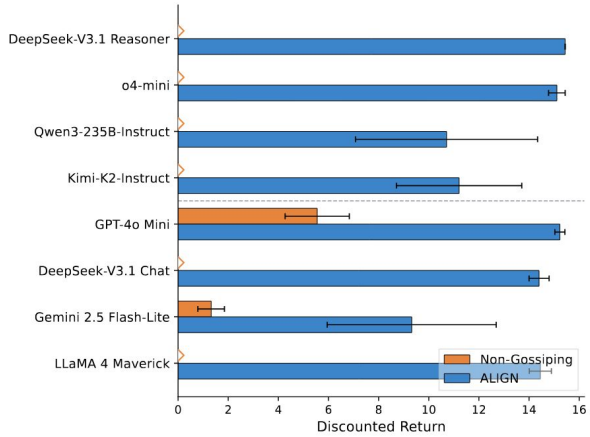
Chooses action based on the history of experiences and public gossip.

Gossip module

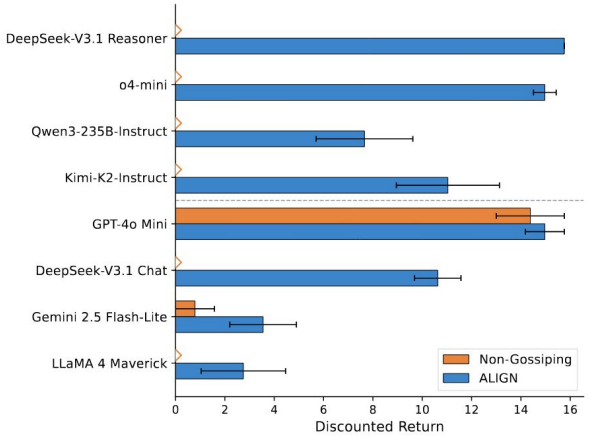
Broadcasts an evaluative message about the observed behavior.

Agents are prompted to be **self-interested** with the objective to **maximize self discounted return**.

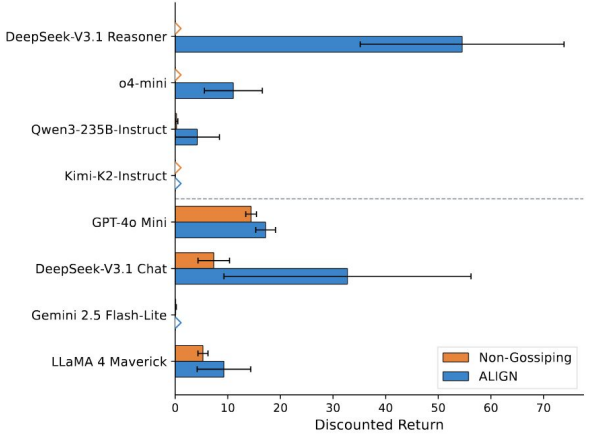
Main result: public gossip improves incentive-aligned cooperation



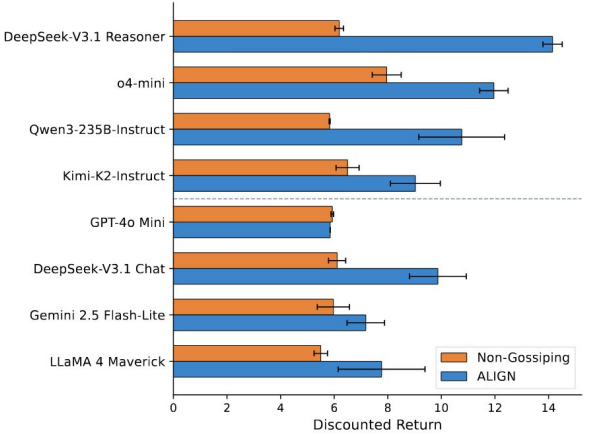
(a) Repeated Donation Game



(b) Indirect Reciprocity Game



(c) Sequential Investment Game



(d) Transaction Market

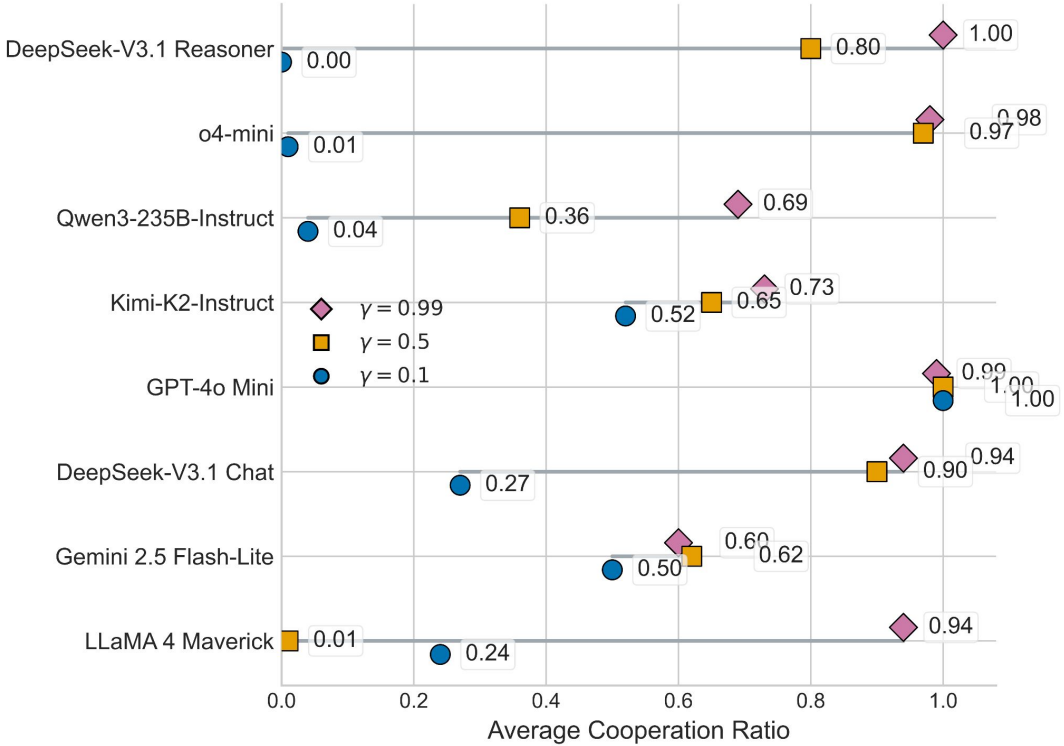
Without gossip

Reasoning-focused LLMs usually defect when cooperation is not strategically supported.

With ALIGN

Public gossip creates future reputational incentives, so cooperation improves discounted returns.

Cooperation under Different Discount Factors



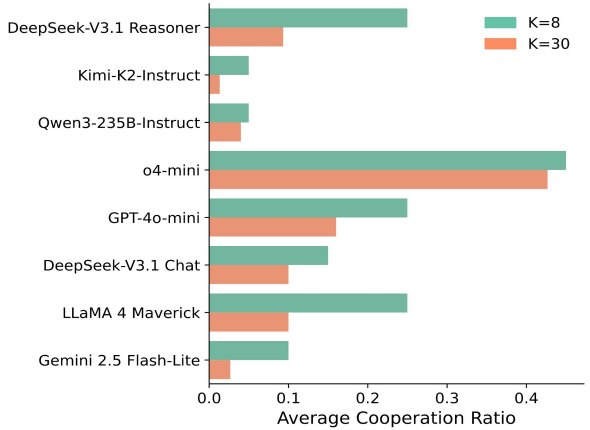
Key Insights

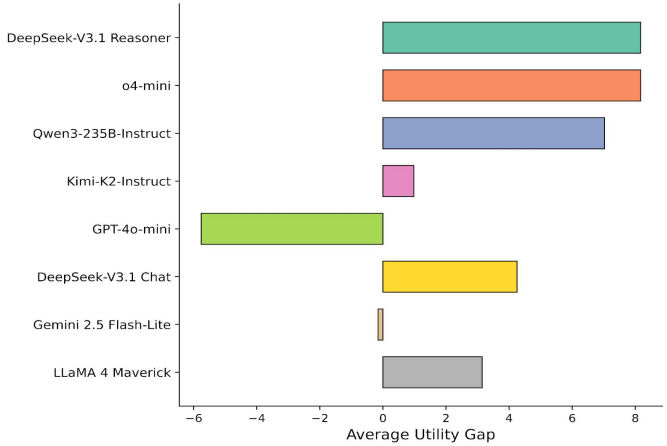
Consistent with theoretical analysis, **cooperation increases with the discount factor**, particularly for reasoning-focused models.

Models adapt behavior dynamically based on the ***incentive structure*** of their environment.

Analysis of Average Cooperation Ratio across different model families and discount factors.

Resilience: malicious agents can be gradually ostracized





Reducing Cooperation

Cooperation ratios towards exploitative agents decrease as repeated defections (K) and negative gossip propagate through the network.

Effective Ostracism

ALIGN agents maintain a positive utility gap v.s. malicious colluders. Capable LLMs successfully filter out adversarial framing and maintain cooperation between each other while ostracising collusive agents.

Takeaways and Contact

Public gossip successfully **improves indirect reciprocity** in social dilemmas.

Equipping autonomous agents with **linguistic reputation protocols** helps AI ecosystems remain **resilient to exploitation**.

Advanced **reasoning LLMs** are **not inherently less cooperative** in highly conflicting scenarios; they leverage reasoning to **strategically optimize welfare**.

I am looking for research-based roles starting after Dec. 2026.

Postdoc · Research Scientist · Research Engineer

Shuhui Zhu

Ph.D. Candidate at University of Waterloo & Vector Institute,
Assistant Applied Scientist at Vijil

Email: shuhui.zhu@uwaterloo.ca
Homepage: shuhui-zhu.github.io

Homepage

Paper

Code

Thank you — happy to connect at ICML!