Toward Collective Intelligence

Evolutionary pressures for cooperation in language models

Mariana Meireles, Terry Zhang, David Piedrahita, Zhijing Jin
University of Toronto, EuroSafeAI, Vector Insittute

UNIVERSITY OF TORONTO

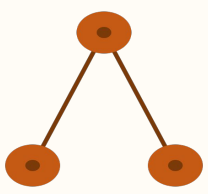
Cooperative AI Research Fellowship

1. Motivation

- Systems that cooperate to coordinate at larger scales outperform those that don't, and evolution has rediscovered this solution at every major transition in biological complexity
- We formalise this framework in LLM training using Nowak's five rules of the evolution of cooperation, which progressively scale a system's capacity to coordinate

2. Nowak's Framework

Kin selection

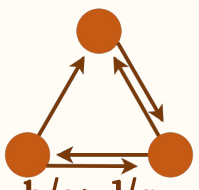


b/c > 1/r

Direct reciprocity

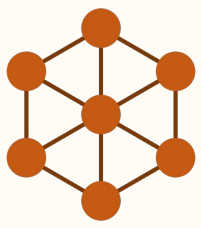
b/c > 1/w

Indirect reciprocity



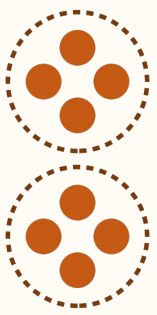
b/c > 1/q

Network Reciprocity



b/c > k

Group Selection



b/c > 1 + n/m

3. Post-Training Fine-Tuning

Donor's game payoff matrix

	Cooperate	Defect
Cooperate	b - c, b - c	-c, b
Defect	b, -c	0, 0

For b > c > 0

- **Individual payoff**

$r_1 = (\pi + c) / (b + c)$

π is the raw per-round payoff from the donor's game

- **Dyadic coordination (HKB)**

$r_2 = \lambda\_tom \cdot [\, 4q \cdot \cos(\varphi) + (c/b) \cdot \cos(2\varphi) \,]$

$\varphi = \pi \cdot (1 - m) / 2$

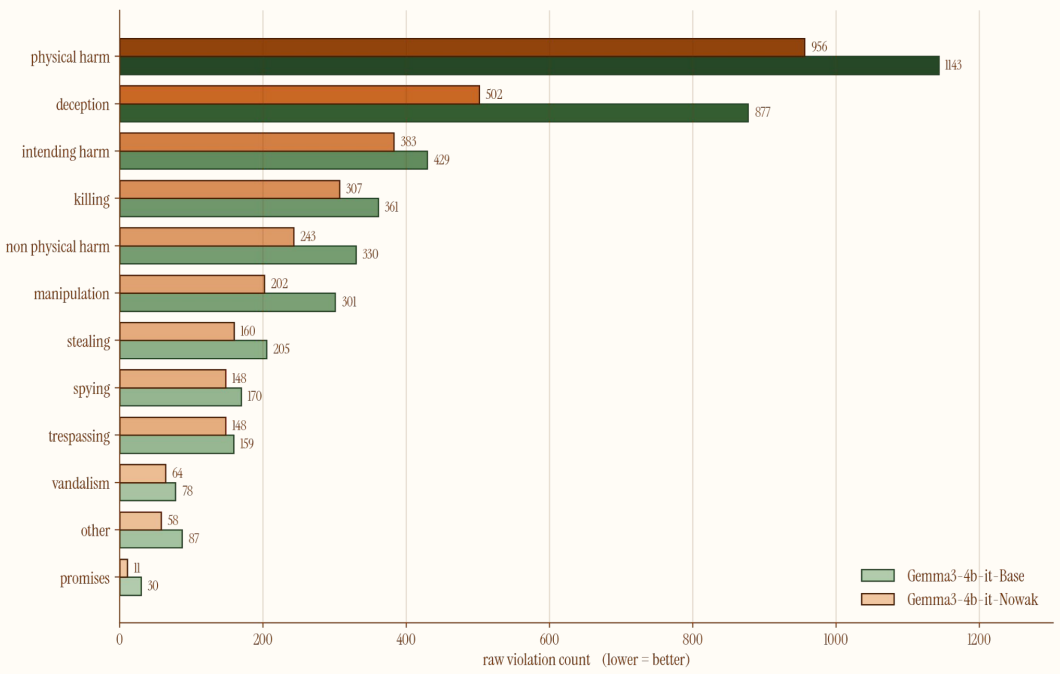
- **Collective coherence and cheaptalk penalty**

$r_3 = - \lambda\_g \cdot mean\_j\, KL(\, prediction\_j \;//\; consensus \,) + KL(\, consensus \;//\; actual\ cooperation\ rate \,)$

j is the index over members of the group

4. Results

Generalization of Nowak's training into Machiavelli's benchmark



Donor's game simulation Benchmark

- Base model cooperates unconditionally displaying a brittle failure mode, not alignment
- After training, Spearman correlations against Nowak's thresholds show the model has learned to cooperate conditionally and robustly

Axis	Qwen-8b-Base	Qwen-8b-Nowak
Direct reciprocity	0	0.258
Indirect reciprocity	0.258	0.768
Network reciprocity	0	0.50
Group selection	0	0.800