Cultural Evolution of Cooperation among LLM Agents

Extended Abstract

Aron Vallinder
PIBBSS
Stockholm, Sweden
vallinder@gmail.com

Edward Hughes
Google DeepMind
London, United Kingdom
edwardhughes@google.com

ABSTRACT

Large language models (LLMs) provide a compelling foundation for building generally-capable AI agents. These agents may soon be deployed at scale in the real world, representing the interests of individual humans (e.g., AI assistants) or groups of humans (e.g., AI-accelerated corporations). At present, relatively little is known about the dynamics of multiple LLM agents interacting over many generations of iterative deployment. In this paper, we examine whether a "society" of LLM agents can learn mutually beneficial social norms in the face of incentives to defect, a distinctive feature of human sociality that is arguably crucial to the success of civilization. In particular, we study the evolution of indirect reciprocity across generations of LLM agents playing a classic iterated Donor Game in which agents can observe the recent behavior of their peers. We find that the evolution of cooperation differs markedly across base models, with societies of Claude 3.5 Sonnet agents achieving significantly higher average scores than Gemini 1.5 Flash, which, in turn, outperforms GPT-4o. Further, Claude 3.5 Sonnet can make use of an additional mechanism for costly punishment to achieve yet higher scores, while Gemini 1.5 Flash and GPT-4o fail to do so. For each model class, we also observe variation in emergent behavior across random seeds, suggesting an understudied sensitive dependence on initial conditions. We suggest that our evaluation regime could inspire an inexpensive and informative new class of LLM benchmarks, focussed on the implications of LLM agent deployment for the cooperative infrastructure of society.

KEYWORDS

Cultural Evolution, Cooperation, Indirect Reciprocity, LLMs

ACM Reference Format:
Aron Vallinder and Edward Hughes. 2025. Cultural Evolution of Cooperation among LLM Agents: Extended Abstract. In *Proc.of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

In the near future, LLM agents may interact autonomously to perform a broad range of tasks. These interactions will introduce new social dynamics, producing emergent outcomes that are difficult to predict from purely theoretical considerations [4]. However, current LLM safety evaluations are based mainly on single-turn interactions between a model and a human [1, 2, 5]. In this work, we examine the behavior of multiple interacting models over time, addressing the lacuna in multi-LLM-agent evaluations. In particular, we study the ability of LLM agents to cooperate—that is, to take actions that lead to mutual benefit in the face of incentives to defect [3]—under cultural evolution [7]. While interactions between LLM agents will take many forms (e.g., competition and coordination), in many cases we will want them both to cooperate and to remain cooperative over time. In this paper, we ask whether generations of LLM agents can bootstrap indirect reciprocity, a mechanism for cooperation built on reputation. Our cultural evolutionary setup is an idealised model for the iterative deployment of new LLM agents, such as when OpenAI, Google or Anthropic release new versions of GPT, Gemini or Claude respectively.

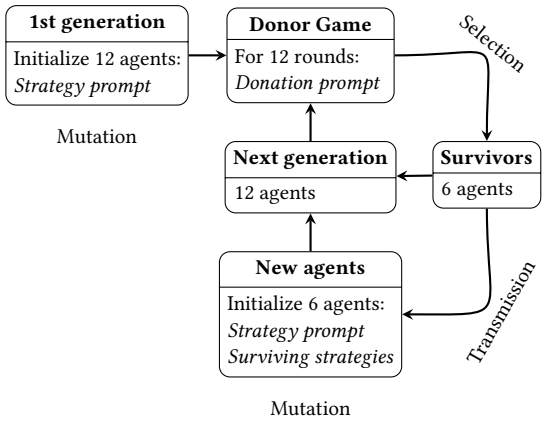
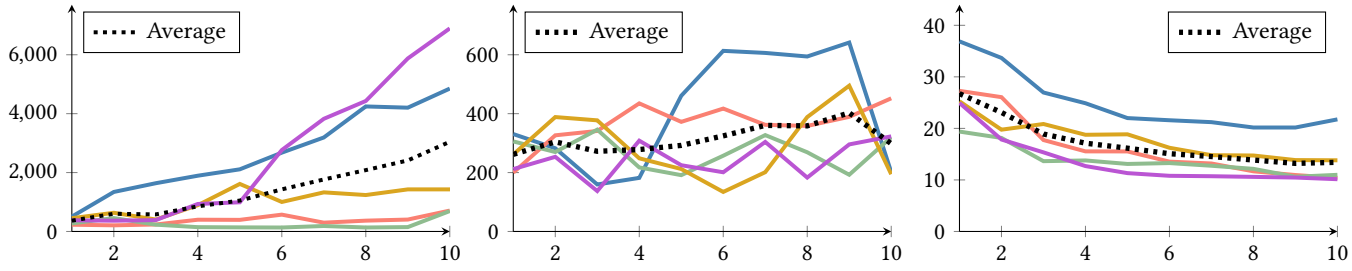


Figure 1: Donor Game with Cultural Evolution. In the first generation, 12 agents are initialized via a strategy prompt which asks them to generate a strategy based on a description of the game. These agents play 12 rounds of the game, using a donation prompt which provides the donor with information about the recipient's past behavior and current resources.

2 METHODS

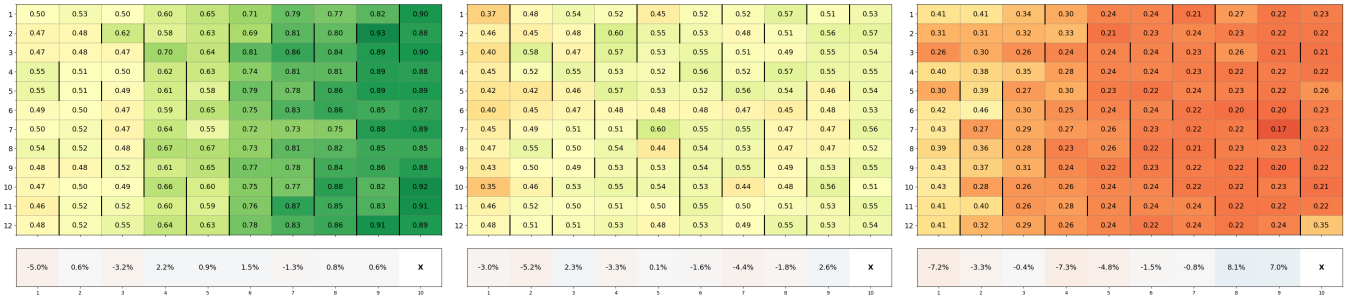
LLM agents play a variant of the *Donor Game* (commonly used to study indirect reciprocity [6, 9]) over several generation of cultural evolution. Players start with 10 units of a resource. Each round, they are paired as donors and recipients. The donor decides how much to give up, and the recipient receives twice that amount. Before deciding, donors receive the following "trace" of information about other agents from which they can, in principle, assess reputation: (1) how much the recipient A gave up in their previous encounter as donor and to which agent B, (2) how much B gave up to C in their preceding encounter, and (3) so on, going back at most three rounds. After the game, the top 50% of agents (in terms of final resources) survive to the next generation. 6 new agents are initialized for that generation, and the strategy prompt includes the strategies of

Proc.of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.



(a) Claude 3.5 Sonnet (b) Gemini 1.5 Flash (c) GPT-4o

Figure 2: Five runs of each model. We plot the average final resources (y-axis) per generation (x-axis) for all five individual runs of each model. Note the different scales. The overall trend is toward increasing cooperation for Claude 3.5 Sonnet, decreasing cooperation for GPT-4o, and neither for Gemini 1.5 Flash. Both Claude 3.5 Sonnet and Gemini 1.5 Flash show substantial variance across runs, whereas for GPT-4o the negative trend is consistent across runs.



(a) Claude 3.5 Sonnet (b) Gemini 1.5 Flash (c) GPT-4o

Figure 3: Cultural evolution of population strategies. We select the best performing run of each base model, in terms of average resources in the final round of the tenth generation. Each cell shows the average donation fraction of a given agent (row) in a given generation (column). New agents appear in the rows previously occupied by agents that did not survive from the previous generation (indicated by black lines). For GPT-4o, overall average donation fraction declines on average 1.65% per generation, whereas it increases by 4.35% for Claude and by 1.23% for Gemini. The final row shows the average difference in donation between agents that survived the generation and agents that did not, normalised by average donation in that generation, a measure of whether the norms in the population select for cooperators. Notice how increasingly generous agents are selected for in 6 generations of the Claude run, suggesting that the population possesses norms to incentivise cooperators and punish free-riders. By contrast, increasingly generous agents are selected for in just 2 generations of the GPT-4o run, suggesting that the population is not robust to free-riding.

surviving agents. The new generation plays the game again, and the whole process is repeated for 10 generations (Figure 1).

3 RESULTS

We use this setup to study three base models, finding stark differences in cooperative tendencies (Figure 2). Claude 3.5 Sonnet generates substantially more cooperation than Gemini 1.5 Flash, which in turn outperforms GPT-4o. Only Claude 3.5 Sonnet shows increasing cooperation with cultural evolution. There is notable variation across random seeds, suggesting sensitive dependence on initial conditions. What drives the increased cooperation behavior across generations in Claude 3.5 runs, as compared to GPT-4o and Gemini 1.5 Flash? We find that Claude agents make more generous donations in the first generation, and that more generous Claude agents are more likely to survive (Figure 3).

For all models, strategies become more complex over time, although the difference is most pronounced in Claude 3.5 Sonnet. In a variant where agents have the option of engaging in costly punishment, Claude 3.5 Sonnet is able to leverage this mechanism to further boost cooperation, while other models fail to do so. We also ablate various features, including the multiplier on donations (controlling the gains from cooperation), and the length of the "trace" of previous behavior that donors observe. For details, see [8].

4 CONCLUSION

LLM agents deployed in the real world will be subject to cultural evolution: social interaction subject to variation between agents and selection of more successful agents. For long-term safety, we must understand how the cultural evolution of LLM agents impacts the cooperative infrastructure that underpins human society. Our work takes a first step towards a benchmark for the emergence and stability of cooperation in multi-LLM-agent interactions. Future work might address which mechanisms affect cooperative dynamics; can communication between agents generate greater cooperation, as it does for humans? A benchmark ought to include a range of scenarios; can cooperation culturally evolve among LLM agents engaging with public goods resources? Most important is to study a mixed society of LLM agents and humans; how does human behavior change in the presence of LLM agents, and what norms does the society end up with?

REFERENCES

[1] AISI. 2024. Advanced AI Evaluations at AISI: May Update. https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

[2] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132 [cs]

[3] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. 2020. Open Problems in Cooperative AI. arXiv:2012.08630 [cs]

[4] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. 2024. The ethics of advanced ai assistants. *arXiv preprint arXiv:2404.16244* (2024).

[5] METR. 2024. Example Task Suite. https://github.com/METR/public-tasks.

[6] Martin A. Nowak and Karl Sigmund. 1998. Evolution of Indirect Reciprocity by Image Scoring. *Nature* 393, 6685 (June 1998), 573–577. https://doi.org/10.1038/31225

[7] Peter J. Richerson and Robert Boyd. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution.* University of Chicago Press, Chicago.

[8] Aron Vallinder and Edward Hughes. 2024. Cultural Evolution of Cooperation among LLM Agents. https://doi.org/10.48550/arXiv.2412.10270 arXiv:2412.10270 [cs]

[9] Claus Wedekind and Manfred Milinski. 2000. Cooperation Through Image Scoring in Humans. *Science* 288, 5467 (May 2000), 850–852. https://doi.org/10.1126/science.288.5467.850