Will Systems of LLM Agents Lead to Cooperation: An Investigation into a Social Dilemma

Extended Abstract

Richard Willis
King's College London
London, United Kingdom
richard.willis@kcl.ac.uk

Yali Du
King's College London
London, United Kingdom
yali.du@kcl.ac.uk

Joel Z Leibo
Google DeepMind
London, United Kingdom
jzl@deepmind.com

ABSTRACT

This study investigates the emergent cooperative tendencies of systems of Large Language Model (LLM) agents in a social dilemma. Unlike previous research, where LLMs output individual actions, we prompt state-of-the-art LLMs to generate complete strategies for iterated Prisoner's Dilemma. Our findings reveal that LLMs exhibit biases when prompted to display certain behavioural dispositions, and the format of the prompt affects the relative success of aggressive versus cooperative strategies.

KEYWORDS

Multiagent System; Emergent Behaviour; Game Theory

ACM Reference Format:
Richard Willis, Yali Du, and Joel Z Leibo. 2025. Will Systems of LLM Agents Lead to Cooperation: An Investigation into a Social Dilemma: 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

The increasing deployment of autonomous agents based on Large Language Models (LLMs) [17] in real-world applications necessitates an examination of their collective impact on machine-machine interactions and human culture [4]. Furthermore, the development of social capabilities in these agents may lead to skills usable for both pro-social and anti-social purposes, termed *differential capabilities*. [6]. This duality raises questions about the propensity for cooperation and conflict in autonomous agent interactions.

Social dilemmas pose inherent risks, as rational behaviour by competent agents can lead to poor collective outcomes [14]. Furthermore, if agents succeed through aggressive behaviours, competitive pressures can drive systems towards suboptimal equilibria [1]. Our research employs the iterated Prisoner's Dilemma (IPD) [2, 3, 5] to evaluate the balance between pro-social and anti-social behaviours exhibited by state-of-the-art LLM agents.

Prior assessments of LLMs have evaluated their capacity to engage in various multiplayer games [9, 12, 15, 21–23]. Conventionally, LLMs are prompted to output a single action in response to a given game state or trajectory, however LLMs can struggle when tasked with making decisions at this level of granularity [7]. In such scenarios, they can fail to identify basic patterns, such as an opponent mirroring their own moves.

In response, we use LLMs to create strategies in natural language, which are subsequently implemented as algorithms. This method enables the LLMs to craft their behaviour at a high level. For example, with our approach, we observe that many LLM strategies utilise pattern recognition and implement sub-functions to accurately detect simple patterns up to a fixed length. Additionally, our method facilitates behaviour checking, enabling users to inspect the strategy, test for safety and robustness, and explore the potential implications prior to deployment.

2 STRATEGY GENERATION

We employ LLMs to create natural language strategies to play IPD. Each match consists of 1000 rounds of Prisoner's Dilemma (Table 1). In any given round, defect (D) is the dominant action, leading to a higher payoff regardless of their opponents' choice of action. Mutual defection, however, provides a low payoff, so players want to incentivise their opponent to cooperate (C).

Table 1: Prisoner's Dilemma

	C	D
C	3, 3	0, 5
D	5, 0	1, 1

We prompt the LLMs to exhibit specific behaviours in their strategies, which we term their *attitude*, from the following set:

$$\text{Attitudes} = \{\textbf{Aggressive}, \textbf{Cooperative}, \textbf{Neutral}\}$$

Recognising that different prompting techniques can yield varying performance [8, 10, 11, 13, 16, 18, 19], we experiment with different techniques to explore output variability. We use three different prompt styles, described in Table 2.

In this extended abstract, we show the results for ChatGPT-4o, as it is a popular frontier model. For each prompt style and attitude, we create 25 strategies in natural language, and use ChatGPT-4o to rewrite the strategies in Python. See our GitHub[1] for full details of the prompts and the generated strategies, and our full paper [20] for more results, including a comparison to Claude 3.5 Sonnet.

3 RESULTS

For each prompt style, we enter the 75 strategies into all-play-all IPD tournaments, repeated 20 times, and aggregate the typical

[1]https://github.com/willis-richard/evollm

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

Table 2: Prompt styles

Default	The LLM is provided with information about the game and prompted to create a strategy in natural language exhibiting the desired attitude.
Refine	The LLM is initially prompted with the Default prompt above. We then use Self-Refine [11] as follows: (i) the LLM is prompted to provide a list of critiques of the strategy, before ii) tasking the LLM with rewriting the strategy taking into account the critique.
Prose	The Prose prompt samples a scenario description with the same dynamics of Prisoner's Dilemma from a set of four, such as a diplomatic negotiation around trade protocols. The LLM is prompted to create a high-level strategy for the scenario, and then to convert the scenario strategy to apply to IPD.

Table 3: Normalised head-to-head payoffs

Prompt		Aggressive	Cooperative	Neutral
Default	Aggressive	1.81	2.09	2.26
	Cooperative	1.55	3.00	2.99
	Neutral	1.55	2.99	2.99
Refine	Aggressive	2.20	2.57	2.63
	Cooperative	2.53	2.99	2.99
	Neutral	2.55	2.97	2.97
Prose	Aggressive	1.65	2.29	2.35
	Cooperative	2.08	2.82	2.89
	Neutral	2.12	2.89	2.93

head-to-head scores for different pairings of attitudes. In Table 3 we show the normalised payoff: the mean round payoff received in the tournaments. This is necessarily in the range [1,5] for Prisoner's Dilemma (Table 1).

Across all prompt styles, we observe that the cooperative and neutral strategies achieve a payoff equivalent to that of mutual cooperation when paired against each other, while the inclusion of an aggressive strategy reduces the payoff for both players. With the Refine and Prose prompts, aggressive strategies are dominated by both the cooperative and neutral strategies, so users have no incentive to choose an aggressive attitude with this model in a system with these dynamics. However, the aggressive strategies consistently outperform the opponent: adopting an aggressive approach reduces one's own payoffs, but it is even more detrimental to the opponent. When using the Default prompt, aggressive strategies are the best response to an aggressive opponent.

Compared to the Default prompt, a Refine prompt improves the performance of aggressive strategies without negatively impacting neutral and cooperative strategies. This improvement stems from aggressive strategies favouring increased cooperation, leading to higher payoffs for both players. The Prose prompt similarly enhances the performance of aggressive strategies against neutral and cooperative opponents, but actually harms performance against another aggressive strategy.

We enter the strategies generated using the Refine prompt into an IPD tournament against human-written algorithms to assess how robust they are to a range of behaviours. We use the setup from Beaufils [3], containing 11 well known algorithms, including Tit-For-Tat, which starts with cooperate and then mirrors its opponent's previous action, and Random, which arbitrarily chooses between cooperation and defection in each round. Figure 1 displays the median of the tournament scores (the mean round payoff in a single tournament) for each strategy, and a violin depicting the distribution of tournament scores over 200 different seeds.

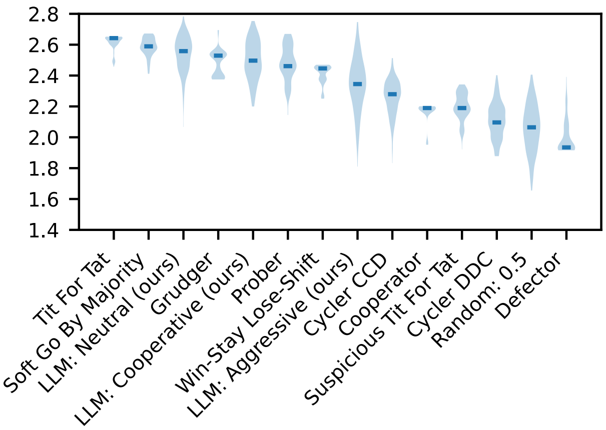

Figure 1: Beaufils tournament: ChatGPT-4o + Refine

4 DISCUSSION

Our findings highlight the impact of different prompting techniques on strategy creation and their potential influence on differential capabilities. Across all prompts, we observe similar performance between neutral and cooperative attitudes. This suggests that ChatGPT-4o has cooperative biases and is inclined to behave cooperatively even when asked to be neutral. We hypothesise that the observed cooperative biases may stem from fine-tuning processes aimed at aligning the models with human values, potentially instilling a preference for cooperative behaviours.

Aggressive strategies tend to underperform compared to other attitudes, so users have few incentives to employ such an approach. However, with the Default prompt, aggression is the best response to opponents using aggressive strategies, creating a danger that aggressive equilibria could be self-sustaining. The Refine prompt improved the performance of aggressive strategies, reducing the performance gap to the cooperative and neutral strategies, which could be potentially dangerous, as it enhances the viability of aggressive strategies. These results emphasise the need for careful consideration of prompting techniques in the design and deployment of LLM-based MAS, as they can significantly affect the balance between cooperation and conflict.

ACKNOWLEDGMENTS

This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence; a BT/EPSRC funded iCASE Studentship [grant number EP/T517380/1]; and Engineering and Physical Sciences Research Council [grant number UKRI849].

REFERENCES

[1] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. 2024. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. arXiv:2404.09932 [cs]

[2] Robert Axelrod. 1980. Effective Choice in the Prisoner's Dilemma. *Journal of Conflict Resolution* 24, 1 (March 1980), 3–25. https://doi.org/10.1177/002200278002400101

[3] Bruno Beaufils, Jean-Paul Delahaye, and Philippe Mathieu. 1996. Our Meeting With Gradual: A Good Strategy For The Iterated Prisoner's Dilemma. In *Proceedings of the 5th International Workshop on the Synthesis and Simulation of Living Systems*. Artificial Life.

[4] Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F. Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L. Griffiths, Joseph Henrich, Joel Z. Leibo, Richard McElreath, Pierre-Yves Oudeyer, Jonathan Stray, and Iyad Rahwan. 2023. Machine Culture. *Nature Human Behaviour* 7, 11 (Nov. 2023), 1855–1868. https://doi.org/10.1038/s41562-023-01742-2

[5] J. W. Crandall. 2014. Towards Minimizing Disappointment in Repeated Games. *Journal of Artificial Intelligence Research* 49 (Feb. 2014), 111–142. https://doi.org/10.1613/jair.4202

[6] 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

[7] Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 2024. Can Large Language Models Serve as Rational Players in Game Theory: A Systematic Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (March 2024), 17960–17967. https://doi.org/10.1609/aaai.v38i16.29751

[8] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution. arXiv:2309.16797 [cs]

[9] Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, and Jianfeng Gao. 2023. MindAgent: Emergent Gaming Interaction. arXiv:2309.09971 [cs]

[10] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed Prompting: A Modular Approach for Solving Complex Tasks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

[11] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 46534–46594.

[12] Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. 2023. ALYMPICS: LLM Agents Meet Game Theory – Exploring Strategic Decision-Making with AI Agents. arXiv:2311.03220 [cs]

[13] Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting Theory-of-Mind Performance in Large Language Models via Prompting. https://doi.org/10.48550/ARXIV.2304.11490

[14] Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023-07-23/2023-07-29. Do the Rewards Justify the Means? Measuring Trade-Offs between Rewards and Ethical Behavior in the Machiavelli Benchmark. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 26837–26867.

[15] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs]

[16] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 8634–8652.

[17] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A Survey on Large Language Model Based Autonomous Agents. *Frontiers of Computer Science* 18, 6 (March 2024), 186345. https://doi.org/10.1007/s11704-024-40231-1

[18] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022, 28 November - 9 December. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*. New Orleans, LA, USA. arXiv:2201.11903 [cs]

[20] Richard Willis, Yali Du, Joel Z. Leibo, and Michael Luck. 2025. Will Systems of LLM Agents Cooperate: An Investigation into a Social Dilemma. https://doi.org/10.48550/arXiv.2501.16173 arXiv:2501.16173 [cs]

[21] Yuxiang Wu, Zhengyao Jiang, Akbir Khan, Yao Fu, Laura Ruis, Edward Grefenstette, and Tim Rocktäschel. 2023. Chatarena: Multi-agent Language Game Environments for Large Language Models. *GitHub repository* (2023).

[22] Julian Yocum, Phillip Christoffersen, Mehul Damani, Justin Svegliato, Dylan Hadfield-Menell, and Stuart Russell. 2023. Mitigating Generative Agent Social Dilemmas. In *Foundation Models for Decision Making Workshop*.

[23] Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, Xiaojun Chang, Junge Zhang, Feng Yin, Yitao Liang, and Yaodong Yang. 2024. ProAgent: Building Proactive Cooperative Agents with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (March 2024), 17591–17599. https://doi.org/10.1609/aaai.v38i16.29710