

Peer Review

Review of: "AI & Cooperation: Replicating the Axelrod Experiment with LLMs"

Burak Ünveren¹

1. Economics, Yildiz Technical University, Turkey

This is a very interesting and creative work. My first concern is that the author trusts the LLMs' judgments regarding their behavioral patterns. In particular, given their responses to the final prompt, the author concludes "Clearly, they are adjusting based on a change in the anticipated choices of the opponent."

However, I typically do not find LLM's introspective comments reliable. Moreover, according to my own personal experience, deterioration of performance in the second round of a certain task is a common feature of LLMs. To see whether this is a crucial factor here, I suggest the author give the baseline and unambiguous prompts with fresh memory in separate sessions. More interesting alternatives can also be contrived.

Second, there seems to be some confusing inconsistency. ChatGPT-5.1 says: "When both players are LLMs, stable cooperation is far more attainable than with humans." Nonetheless, the author remarks: "Cooperation declines substantially when LLMs know they are competing against other LLMs."

A clarification would improve the text. It seems that ChatGPT-5.1 is making an erroneous assessment. I already provided an explanation above: introspective comments of LLMs are unreliable. Or maybe I am missing a point, and there is a better explanation.

Despite my concerns above, this is high-quality research. And I really enjoyed reading this innovative work. Congratulations.

Declarations

Potential competing interests: No potential competing interests to declare.