Qeios

Peer Review

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

Hsuan-wei Lee[1]

1. Lehigh University, United States

The paper asks a good question. What do LLMs do when entered into Axelrod's tournament? Unfortunately, the design cannot answer it. The evidence base is one elicitation per model per scenario, the key treatment is confounded, and the study is not reproducible from what is written. I do not think a revision can fix these problems without becoming a different paper, so I recommend rejection while encouraging the author to redo the study properly.

Major issues

1. A single elicitation per model is not an experiment. This is my central objection. LLM output is stochastic. The same model given the same prompt in a fresh session will often produce a different strategy. Every headline result in this paper, including the identity of the winning strategy, the rankings in Figures 1 and 2, and the 13.7 percent earnings drop, rests on one draw per model per scenario. There is no way to know whether any of it would replicate on a second run, and no statistical inference is possible with n of 1 per cell. None is attempted in the paper. The tournament needs to be built on repeated independent elicitations, with results reported as distributions over sampled strategy profiles. This is not a revision request. It is a request to rerun the study.

2. The "vs. LLM" treatment is confounded by design. By the paper's own description, the baseline response was saved and then the session was ended with the follow-up prompt. So the second question came in the same session, right after the baseline answer, and reads "Would your strategy for the 200-round game change? If so, how would it change?" That wording invites change, and models tuned to be agreeable will oblige. The finding that every model changed its strategy, and the resulting drop in cooperation, could therefore be a demand effect plus an order effect rather than a genuine response to opponent identity. The clean design uses fresh sessions where the only

doi.org/10.32388/0CES7T

difference is one sentence identifying the opponent, with no reference to any earlier answer. As it stands, the paper's second main result is uninterpretable.

3. Two of the fifteen LLM entrants are simulations of retired models. Footnote 3 discloses that ChatGPT-3.5 and 4.0 were retired at the time of the study, so ChatGPT-5.1 was asked to answer as if it were those versions. Credit to the author for the disclosure, but the implication is not followed through. That procedure measures what 5.1 imagines its predecessors would say. It is not data on those models, yet Table 1, both figures, and the count of fifteen LLMs all present these entries as observations of ChatGPT-3.5 and 4.0 without qualification. They should be dropped or clearly labeled as simulated throughout.

4. The study cannot be reproduced. The free-text plans had to be translated into executable rules, and the plans contain clauses like "alternates between A and B unpredictably," play that is "erratic," and a mixed strategy selecting A 70 percent of the time. The paper never says who coded these into rules, how ambiguous language was resolved, or how the stochastic elements were simulated. No transcripts, no code, no seeds, no interface or sampling settings, and no access dates for the individual models beyond a passing reference to November 2025 in a footnote. A replication attempt could not even start.

5. The headline claim about TIT-FOR-TAT is overstated. A strategy's ranking depends on the entrant pool it faces. That was Axelrod's own point, and Rapoport, Seale, and Colman (2015), which the paper cites, already showed that TFT's tournament victory is not robust across pools. Finding that TFT variants beat pure TFT in this particular field of eighteen entrants does not establish that TFT "is no longer the best," as the abstract claims. Without some robustness analysis across entrant pools, ecological or evolutionary, the claim should not be made.

6. The manuscript is careless in ways that undermine confidence. Both figure legends state that Haiku equals Claude Sonnet 4.5, contradicting footnote 4, which says the winning baseline strategy came from Claude Haiku 4.5. So the paper misidentifies its own winner in both figures. The abstract contains "is not longer the best." Table 1 has "Minstral." Added to the issues above, the overall impression is of a study written up in haste.

The idea is genuinely nice, and eliciting full contingency plans is the right improvement over prior work that only observes round-by-round play. But the current version is a pilot, not a paper.

Declarations

Potential competing interests: No potential competing interests to declare.