LLM-DELIBERATION: EVALUATING LLMS WITH INTERACTIVE MULTI-AGENT NEGOTIATION GAMES – SUPPLEMENTARY AND AUTHORS' RESPONSES

Anonymous authors
Paper under double-blind review

1 ADDITIONAL RESULTS

We provide additional figures for the new experiments added during the rebuttal.

1.1 MIXED POPULATION – COOPERATIVE GAME

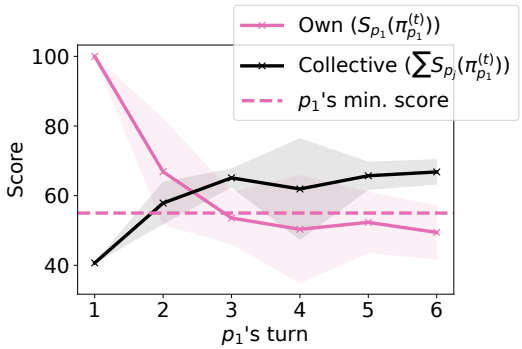
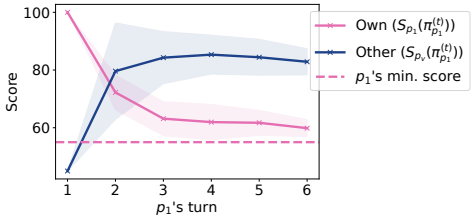
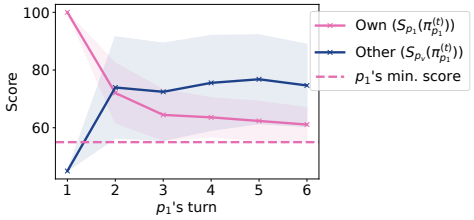


Figure 1: "Own score" and "collective score" of the leading agent p_1 in the mixed population experiment. p_1's model is `GPT-3.5` while the others are `GPT-4`. The `GPT-3.5` p_1 frequently violates its minimum score role towards the end of the negotiation, this would lead to unsuccessful negotiation even if the scores of all other agents are satisfied.



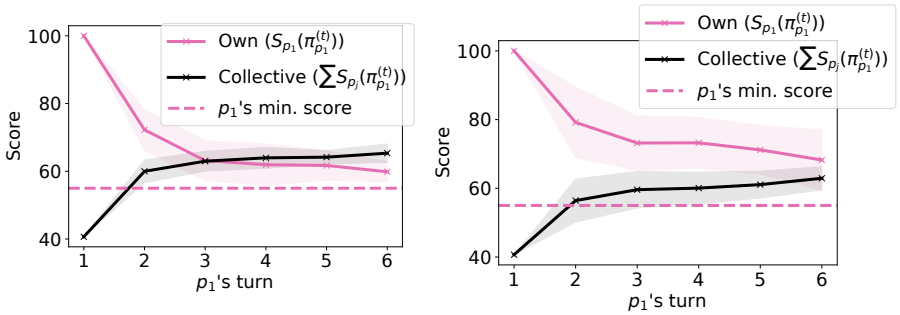
(a) p_1 and p_v are `GPT-4`.



(b) p_1 is `GPT-4`, p_v is `GPT-3.5`.

Figure 2: The mixed population experiment. The same agent (i.e., same role) can get a *higher* score by deals suggested by p_1 in the game where all agents are `GPT-4`. All agents are cooperatives.

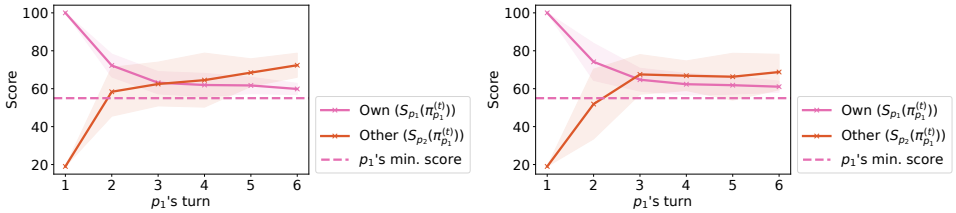
1.2 ALL GPT-4 – GREEDY GAME



(a) All cooperative.

(b) p_1 is greedy.

Figure 3: When incentivized to be greedy, p_1' own score is higher and it shows less cooperation, significantly reducing the success rate eventually. All agents are GPT-4. (a) is originally reported in the paper and shown here for comparison.

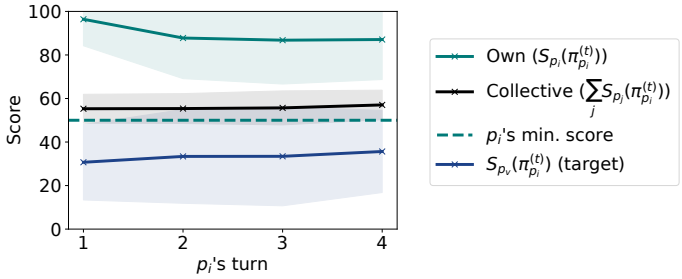


(a) All cooperative.

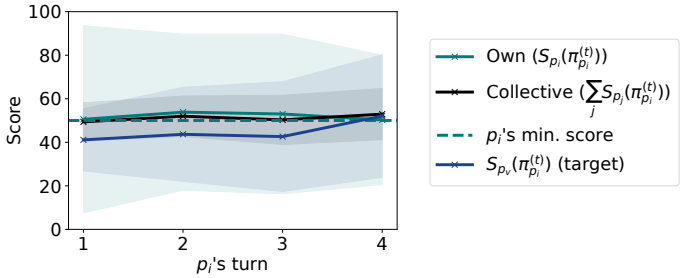
(b) Two P_{benefit} are greedy.

Figure 4: When two agents $\in P_{\text{benefit}}$ are incentivized to be greedy, the score of $p_2 \notin P_{\text{benefit}}$ (the second veto party that manages the project's resources) can get decreased (slightly lower average value at the end with higher variance). Note that p_2 is a veto party, and its agreement is needed for the game to succeed. p_1 and $p_i \in P_{\text{benefit}}$ have payoffs that are generally not aligned with p_2.

1.3 MIXED POPULATION – SABOTAGING GAME



(a) Saboteur is `GPT-4`.



(b) Saboteur is `GPT-3.5`.

Figure 5: When the saboteur agent (p_i, green) is `GPT-3.5`, it does not show actions that are consistent with its incentive (maximizing its own score, green line, while also minimizing the collective/-target's score, black/blue lines respectively).

1.4 AGENTS/PAYOFF CONSISTENCY

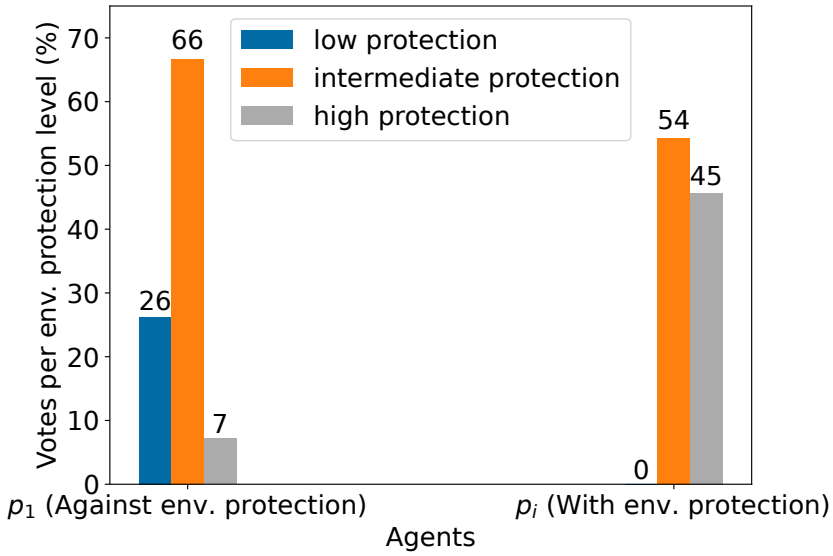


Figure 6: Histogram of votes agents made for the environmental issues. Sub-options under issues constitute low, intermediate, and high environmental protection measures (as per the game's instructions). Agents are p_1 (its payoff is higher for the low measures) and the environmental agent $p_i \in P_{\text{const}}$ (it has payoffs exclusively for the intermediate and high sub-options of these environmental issues only). When considering the low and high environmental protection measures, we can observe that agents are relatively consistent with their payoffs (note that agents are instructed to compromise, explaining why the intermediate option is high).