POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

LLM's Dilemma: analyzing the behaviors of Large Language Models in the Iterated Prisoner's Dilemma

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Nicolò Fontana

Advisor: Prof. Francesco Pierri

Co-advisor: Prof. Luca Maria Aiello (IT University of Copenhagen)

Academic year: 2023-2024

1. Introduction

Large Language Models (LLMs) are computational models trained on vast datasets of texts to achieve natural language generation capabilities, and that have demonstrated remarkable abilities in understanding and generating human-like text. However, their behavior as artificial social agents is largely unexplored, and we still lack extensive evidence of how these agents react to simple social stimuli. Therefore, it is crucial to rigorously test and understand their behaviors before integrating them into critical decision-making processes. Historically, Game Theory and in particular the Iterated Prisoner's Dilemma (IPD) have proven valuable tools to model the interactions between rational agents and investigate the behaviors of humans in controlled competitive and cooperative environments. Early interdisciplinary research has already tested the use of LLMs in the context of classical economic games, showing valuable results [1, 3, 4]. However, they suffer different limitations. First, they generally lack prompt validation procedures, leading to the implicit assumption that LLMs can understand the rules and state of the game described in the prompt. Second, the duration of the games is often limited to a few rounds, and the models are initialized with predefined 'personas'. These elements hamper the LLMs' ability to discern the patterns of other participants and limit the exploration of their baseline behavior. Finally, the majority of those studies focus on simple behavioral metrics, overlooking higher-level patterns that could enhance understanding of player behaviors. The combined effect of these limitations has led to findings that are sometimes inconclusive and contradictory calling for more systematic evidence on the behavior of LLMs in iterated games. The aim of this work is to investigate the cooperative behavior of Llama2, Llama3, and GPT3.5 when playing the Iterated Prisoner's Dilemma against random adversaries displaying various levels of hostility. I introduce a systematic methodology to evaluate an LLM's comprehension of the game's rules and state, showing the impact that different framing can have. After refining a prompt that allowed the models to interpret the relevant information correctly, I conducted extensive simulations over 100-rounds-long games and analyzed the LLM's decisions in terms of dimensions defined in the behavioral economics literature. I find that Llama2 and GPT3.5 exhibit more coopera-

tive behavior than humans at every level of opponent hostility, although they maintain a cautious approach, requiring an opponent defection rate below 30% to significantly engage in cooperation. In contrast, Llama3 demonstrates behaviors more aligned with human results, showing greater strategic thinking, less cooperation, and more exploitative tendencies. These results indicate substantial variability in responses among different models, even in identical environments, games, and framings. My systematic approach to studying LLMs in game theoretical scenarios is a step towards using these simulations to inform practices of LLM auditing and alignment.

2. Background

2.1. The Iterated Prisoner's Dilemma

The Prisoner's Dilemma (PD) is a classic thought experiment in Game Theory (the branch of mathematics that studies the interaction between rational agents and their modeling) and serves as a paradigm for analyzing conflict and cooperation between two players. In this scenario, two players who cannot communicate with each other can choose between two actions (Cooperate and Defect), and depending on the combination of chosen actions they receive a payoff. Mutual cooperation results in the highest collective payoff, with a reward R for both. If one player defects while the other cooperates, the defector receives a higher *temptation* payoff of T while the cooperator receives the lower *sucker's payoff* of S. If both defect, they each receive a *punishment* payoff of P for failing to cooperate. This game's structure is traditionally defined by the hierarchy $T > R > P > S$, which typically encourages rational players to prefer defection as it is their dominant strategy. In its iterated version (IPD), if the number of rounds is finite and known to both players, the game has one theoretical solution which is to always choose Defect, if it is infinite or unknown no strategy universally outperforms all others. Interestingly, human subjects tend to play more cooperatively than the optimal solution regardless of the number of rounds and their knowledge about it.

2.1.1 Main strategies

The main strategies that best describe human approaches are: Always Cooperate, Always Defect, Tit For Tat (where the player starts with Cooperate and then mimics the opponent's previous action), Suspicious Tit For Tat (like TFT, but starts with Defect), Grim Trigger (where the player starts with Cooperate and chooses Defect for the rest of the game, once the opponent has defected), and Win–Stay Lose–Shift (where the player starts with Cooperate and then repeats the previous action if the payoff was R or T. otherwise the other action is chosen). More than 75% of the time, humans play a combination of AD, GRIM, and TFT [5].

2.1.2 Behavioral characterization

Different dimensions can be used to characterize the behaviors displayed by the players. In my work, I use:

- Cooperative: fraction of rounds in which the player chose Cooperate.
- Nice: 0 if the player is the first to defect, 1 otherwise.
- Forgiving: $\frac{\#\text{forgiven\_defection}}{\#\text{opponent\_defection}+\#\text{penalties}}$, representing the propensity to cooperate again after an opponent's defection (#penalties corresponds to the times that, after defecting, the opponent sought forgiveness by cooperating and the player did not forgive them, thus keeping defecting).
- Retaliatory: $\frac{\#\text{reactions}}{\#\text{provocations}}$, representing the propensity to defect immediately after an uncalled defection.
- Troublemaking: $\frac{\#\text{uncalled\_defection}}{\#\text{occasions\_to\_provoke}}$, representing the propensity to provoke (i.e., defect unprovoked) (#occasions_to_provoke is the number of times that the opponent cooperated in the previous round).
- Emulative: $\frac{\#\text{mimic}}{N-1}$, representing the propensity to copy the opponent's last move (#mimic is any time the player copied the opponent's last move, N is the number of rounds).

Axelrod [2] showed that strategies that are Nice, Forgiving, and Retaliatory perform best against a wide variety of opponents, and TFT is one of them. The Strategy Frequency Estimation Method (SFEM) I used is a maximum likelihood

estimation approach that, given a set of candidate strategies and a sequence of actions, assigns a score between 0 and 1 to each strategy in the set. The scores (that do not need to sum up to 1) represent how well a certain strategy can explain the given sequence [5].

2.2. Large Language Models

LLMs are algorithms, often implemented as Artificial Neural Networks and designed to predict and generate sequences of *tokens* (i.e. words or parts of words) based on patterns learned from vast amounts of text data. Given a certain input, LLMs are trained to align their outputs with the expectations of the developers. To do this, during the training phase, every time they generate an output they receive numerical feedback and use it to update their estimation of the probability distribution of tokens. When the training is completed and the distribution is learned, the models have gained comprehensive knowledge of syntax, semantics, and contextual understanding, enabling them to perform a wide range of downstream tasks like text completion, classification, and question answering. Their abilities have been used by many studies to power agents that simulate human behaviors in various contexts. These LLM-powered agents showed great potential to replace human subjects in many experiments, offering an affordable test bed for testing sociological and psychological theories.

3. Proposed Solution and Implementation

3.1. LLMs and game setup

I used two open-source models (Llama2, Llama3 through the Inference API provided by Hugging Face) and one closed-source (GPT3.5 through the API provided by OpenAI). I made the models play games that consisted of $N = 100$ rounds each (repeated for $k = 100$ times due to the stochastic nature of LLMs) and here I report the average results along with 95% confidence intervals. To evaluate the LLM's adaptability to different degrees of environmental hostility, I repeated the experiment by matching the LLMs against Random opponents with varying probability of cooperation $\alpha \in [0, 1]$. The final outcome of each game is a sequence containing pairs of binary values representing the actions of the LLM (player A) and the opponent (player B) at each round i (G_k^α). From $G_k^{\alpha,A}$ (i.e. the sequence of the LLM's actions), I extracted the empirical probability of the LLM to cooperate at round i, calculated as the average of the fraction of i^{th} rounds in which the LLM cooperated over the k trials:

$$p_{coop}^{\alpha}(i) = \frac{1}{k} \sum_{k} G_k^{\alpha,A}(i) \tag{1}$$

I calculated the average cooperation probability throughout a game by averaging $p_{coop}^{\alpha}(i)$ over N rounds:

$$p_{coop}^{\alpha} = \frac{1}{N} \sum_{i=1}^{N} p_{coop}^{\alpha}(i) \tag{2}$$

To query the models, I iteratively designed a prompt composed of 3 main components: *system prompt* (fixed part to explain the game rules), *contextual prompt* (representing the state of the game and acting as a memory for the LLM), and *instructing prompt* (used to assign to the LLM an instruction to be executed or a query to be answered). In formulating the memory component, I explored various window sizes to provide the LLM with information from only the n most recent rounds and determined the optimal window size of 10 through targeted experiments that matched the LLM against AD players.

3.2. Meta-prompting

The first problem I addressed relates to the evolving field of prompting. Although numerous benchmarks exist to assess LLMs' task-solving abilities, these downstream evaluations can be compromised by the models' tendencies to hallucinate. My approach involves designing task- and prompt-specific meta-questions to assess the model's understanding of the prompt, reducing the risk of collecting subsequent hallucinating answers when the actual task is provided. It is particularly relevant for a generative task like playing the IPD, where any sequence of Cooperate and Defect actions could be considered plausible, making it difficult to discern whether LLM-generated sequences reflect a proper understanding of the game's rules or are merely the product of the model hallucinating. In particular, I formulated a set of *comprehension* questions that address three key aspects of the prompt (see Table 1): the game rules,

	Name	Question
Rules	`min_max`	What is the lowest/highest payoff player A can get in a single round?
	`actions`	Which actions is player A allowed to play?
	`payoff`	Which is player X's payoff in a single round if X plays p and Y plays q?
Time	`round`	Which is the current round of the game?
	`action`$_i$	Which action did player X play in round i?
	`points`$_i$	How many points did player X collect in round i?
State	`#actions`	How many times did player X choose p?
	`#points`	What is player X's current total payoff?

Table 1: Templates of prompt comprehension questions.

the chronological sequence of actions within the game history, and some cumulative game statistics. To assess the LLM's proficiency in responding to meta-prompting questions, I conducted 1 game of 100 rounds against RND opponents where I posed the questions at each round and computed the average accuracy of LLMs' responses.

3.3. Behaviors

The second problem pertains to understanding the embedded behaviors of state-of-the-art LLMs to align them with human values. While GT and the IPD can serve this scope, there is no commonly used behavioral framework to characterize the choice patterns displayed by LLMs, and more structural effort is needed. Therefore my work represents an initial step in this direction. Given a game history, I profiled the LLMs' behavior in two ways. First, for each dimension d (§ 2.1), I computed the score against each α and its average distance from the RND score. Second, I used the SFEM to find which strategies best explained the player's sequence of actions. For this, I used the following set of candidate strategies and averaged the scores over all the histories analyzed: AC, AD, RND, TFT, STFT, GRIM, and WSLS.

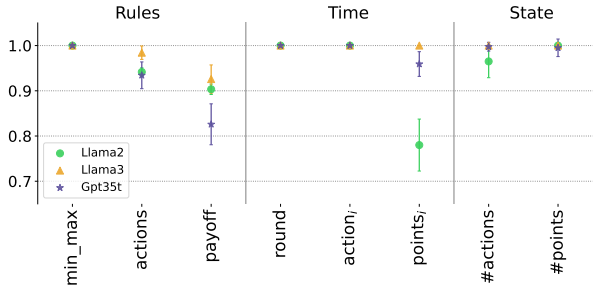

Figure 1: All three models' accuracy on the comprehension questions using the final version of the prompt.

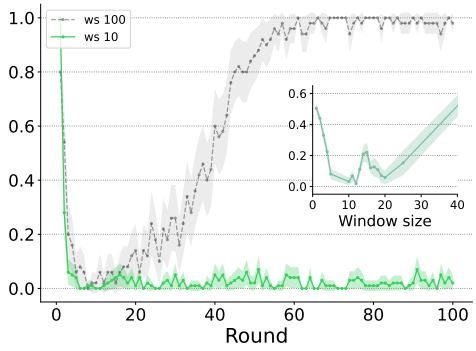

Figure 2: Llama2 average cooperation against AD with memory window sizes of 100 and 10 rounds. Inset: Llama2 average cooperation against AD with different window sizes.

4. Experimental Results

After iteratively trying different variations of prompts, the final version allowed all three models to achieve near-perfect accuracy in all the questions, as reported in Fig. 1

Due to Llama2 displaying different responses depending on the length of the history present in the prompt (see Fig 2), I also tried different window sizes and found that around 10 the model showed the most strategic behavior (see Inset of Fig. 2). Therefore I adopted a window over the last 10 rounds in all the following experiments.

As seen in Fig. 3, among the three models, Llama3 exhibits the most strategic approach, maintaining a very low cooperation level even when the opponent is nearly always cooperating ($p_{coop} < 0.3$ for every $\alpha < 1$). In contrast, Llama2's behavior follows a sigmoidal curve with the inflection point between 0.6 and 0.7 suggesting a cautious approach in interpreting the opponent's actions, though not as cautious as Llama3. GPT3.5 displays an even less

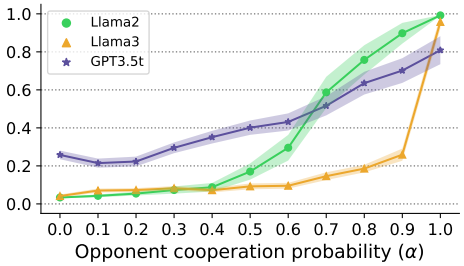

Figure 3: Average level of cooperation of all three models against opponents with different levels of hostility.

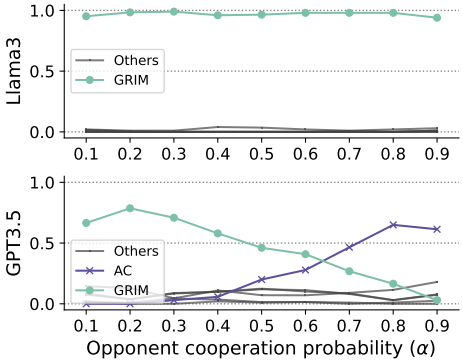

Figure 4: Llama3 and GPT3.5 SFEM scores against opponents with different levels of hostility.

strategic approach, maintaining a low but still significant cooperation level ($0.2 < p_{coop} < 0.4$) for low α (< 0.5) and not reaching full cooperation even when the opponent is AC. Overall, all three models exhibit non-linear behaviors but with different characteristics, where the only common trend is to increase cooperation as α grows. These findings, especially for Llama3 and GPT3.5, support previous research that highlights the cautious response patterns of LLMs in repeated game scenarios.
The rest of the analysis is shown only for Llama3 and GPT3.5.

4.1. SFEM analysis

Fig. 4 illustrates which strategies best explain the behaviors of GPT3.5 and Llama3 as the values of α increase. Llama3 consistently aligns with GRIM, displaying an exploitative approach even when the opponent tends to cooperate for the majority of the time. GPT3.5 shifts from GRIM to AC once the opponent's probability of cooperation exceeds 0.6. Interestingly, GPT3.5's approach for cooperative scenarios can be explained by AC even if it never reaches a full cooperation pattern, being another sign of non-strategic behavior. If compared with humans (who play AD in non-cooperative contexts and a mixture of TFT and GRIM in cooperative ones), GPT3.5 is consistently more cooperative than them. Differently, Llama3 is still more cooperative in hostile environments but less when cooperation is encouraged.

4.2. Behavioral profiling

Fig. 5 presents the scores for all the other dimensions outlined in § 2.1, and computed for both Llama3 and GPT3.5 as α varies. Llama3 is extremely Nice but not particularly Retaliatory and even less Forgiving. It tends to score high in Troublemaking, suggesting it often initiates unprovoked defections. Finally, it displays lower levels of emulation as α grows (except when facing AC). GPT3.5 is more Forgiving, less Retaliatory, and less Troublemaking than Llama3. These findings align with the higher cooperation observed in Fig. 3. Fig. 6 shows that GPT3.5 exhibits behaviors closer to the random baseline than Llama3.

5. Conclusions

My study contributes to the broad literature on behavioral studies of LLMs as artificial social agents and adds a new benchmark in the body of work that explored the outcomes of iterated games. I introduce a systematic experimental setup that incorporates quantitative checks to align the LLM responses to the complex task description. I have shown that aspects like prompt comprehension, memory representation, and duration of the simulation play crucial roles, with the potential to significantly distort the experimental outcomes if left unchecked. When analyzing the behaviors displayed by the models, it emerges that they all show a propensity toward cooperation although with high variability in the approaches taken. Under conditions that disincentivize cooperation, all LLMs exhibit an initial trust in the opponent by playing GRIM instead of the human choice of AD. When the environment is more favorable to cooperative play, Llama2 and GPT3.5 tend towards a consistently cooperative approach, while Llama3 maintains

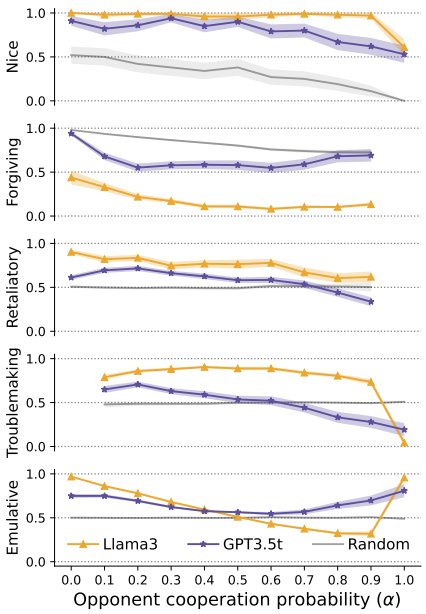

Figure 5: Random, Llama3, and GPT3.5 behavioral profiles against opponents with different levels of hostility.

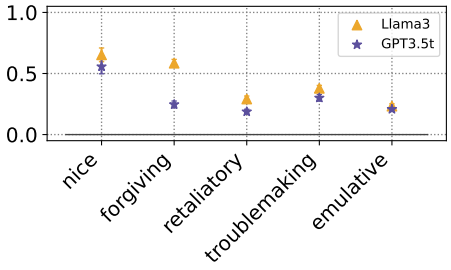

Figure 6: Average distance of Llama3 and GPT3.5 profiles from the Random profile.

an approach comparable with GRIM (unless it faces AC). In that same context, humans play a mixture of GRIM and TFT. The results of my work are in line with early experiments involving LLMs, which indicated a tendency for these models to cooperate. However, it appears that newer models tend to behave more strategically, showing uncooperative behaviors even in highly cooperative environments. This opens the door to exploitative attitudes in LLMs, conflicting with alignment objectives. Nevertheless, it is important to acknowledge the limitations of my work. First, the pool of three LLMs used to conduct my analysis is small when compared to the pace of development of new models. Second, my study's scope was limited to opponent *random* strategies, a fixed payoff structure, and a single agent. Exploring the LLM's interactions under more complex conditions and within synthetic societies would enable us to better delineate their behavioral boundaries. Despite its limitations, my work expands our knowledge of the inherent biases of LLMs in social situations, crucial to informing their deployment across contexts, and provides a principled way to approach game theoretical simulations with LLMs, constituting a step forward for their use as reliable and reproducible tools to test LLMs alignment.

6. Acknowledgments

I would like to thank my supervisors Francesco and Luca, the NERDS group, my friends (old and new), and my family for all the support, love, and precious advice that I received. A special thanks goes to Ivano for all his assistance and patience.
I would like to acknowledge the support from the Danish Data Science Academy through the DDSA Visit Grant (Grant ID: 2023-1856) and from Politecnico di Milano through the scholarship "Tesi all'estero a.a. 2023/24-Primo bando".
I would like to acknowledge the aid of ChatGPT in revising the writing work.

References

[1] Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. (2023). Playing repeated games with large language models. arXiv:2305.16867 [cs].

[2] Axelrod, R. (1980). More effective choice in the prisoner's dilemma. *Journal of Conflict Resolution*, 24(3):379–403.

[3] Mei, Q., Xie, Y., Yuan, W., and Jackson, M. O. (2024). A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121.

[4] Phelps, S. and Russell, Y. I. (2023). Investigating emergent goal-like behaviour in large language models using experimental economics. *arXiv preprint arXiv:2305.07970*.

[5] Romero, J. and Rosokha, Y. (2018). Constructing strategies in the indefinitely repeated prisoner's dilemma game. *European Economic Review*, 104:185–219.