Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Nicer than Humans: How Do Large Language Models Behave in the Prisoner's Dilemma?

Nicoló Fontana[1], **Francesco Pierri**[1], **Luca Maria Aiello**[2,3]
[1]Politecnico di Milano, Italy
[2]IT University of Copenhagen, Denmark
[3]Pioneer Centre for AI, Denmark
nicolo.fontana@mail.polimi.it, francesco.pierri@polimi.it, luai@itu.dk

Abstract

The behavior of Large Language Models (LLMs) as artificial social agents is largely unexplored, and we still lack extensive evidence of how these agents react to simple social stimuli. Testing the behavior of AI agents in classic Game Theory experiments provides a promising theoretical framework for evaluating the norms and values of these agents in archetypal social situations. In this work, we investigate the cooperative behavior of three LLMs (Llama2, Llama3, and GPT3.5) when playing the Iterated Prisoner's Dilemma against random adversaries displaying various levels of hostility. We introduce a systematic methodology to evaluate an LLM's comprehension of the game rules and its capability to parse historical gameplay logs for decision-making. We conducted simulations of games lasting for 100 rounds and analyzed the LLMs' decisions in terms of dimensions defined in the behavioral economics literature. We find that all models tend not to initiate defection but act cautiously, favoring cooperation over defection only when the opponent's defection rate is low. Overall, LLMs behave at least as cooperatively as the typical human player, although our results indicate some substantial differences among models. In particular, Llama2 and GPT3.5 are more cooperative than humans, and especially forgiving and non-retaliatory for opponent defection rates below 30%. More similar to humans, Llama3 exhibits consistently uncooperative and exploitative behavior unless the opponent always cooperates. Our systematic approach to the study of LLMs in game theoretical scenarios is a step towards using these simulations to inform practices of LLM auditing and alignment.

1 Introduction

Large Language Models (LLMs) can operate as social agents capable of complex, human-like interactions (Park et al. 2023). Their integration into online social platforms is unfolding rapidly (Cao et al. 2025; Yang and Menczer 2024), presenting potentially severe risks (Floridi and Chiriatti 2020; Ferrara 2024; Nogara et al. 2025)—especially due to their biases (Liu, Bono, and Pierri 2025; Saffari et al. 2025)—as well as intriguing opportunities (Dafoe et al. 2020; Breum et al. 2024), for example, as fact-checking tools (Fontana et al. 2025). To understand and anticipate the behavioral dynamics that may arise from the interaction between artificial agents and humans, it is essential to first study how these agents react to simple social stimuli (Bail 2024). Behavioral economics experiments, particularly those grounded in Game Theory, provide an ideal ground for testing the responses of AI agents to archetypal social situations (Horton 2023). These experiments typically involve goal-oriented scenarios where multiple 'players' engage in a series of repeated interactions (Osborne and Rubinstein 1994). To optimize for the goal, the decisions taken at each round must strategically account for the anticipated actions of the other players. However, the decisions of human participants often deviate from the theoretically optimal strategies due to the influence of social and psychological factors that conflict with the game's objectives (Camerer 1997). Similarly, given that LLMs are computational models built upon collective human knowledge and culture (Schramowski et al. 2022), observing their behavior in classic iterated games could shed light on the social norms and values that these models reflect, as well as their capability in reasoning, planning, and collaborating in social settings.

Early interdisciplinary research has explored the use of LLMs within the context of classical economic games (see Related Work). While highly valuable, all these studies exhibit at least one of the following limitations. First, they generally lack prompt validation procedures, leading to an implicit assumption that LLMs can understand the complex rules of the game and the history of past actions described in the prompt (Akata et al. 2023; Mao et al. 2025; Mei et al. 2024). Second, the duration of simulated games is often limited to a few rounds (Akata et al. 2023; Brookins and DeBacker 2024; Fan et al. 2024; Guo 2023; Xu et al. 2024), hampering the LLMs' ability to discern the decision-making patterns of other participants—a phenomenon we show in our own experiments. Third, the initialization of LLMs with predefined 'personas' tends to skew their responses towards pre-determined behaviors, such as altruism or selfishness (Brookins and DeBacker 2024; Fan et al. 2024; Guo 2023; Horton 2023; Lorè and Heydari 2024; Phelps and Russell 2023). This approach limits the exploration of the LLMs' baseline behavior, which is crucial for understanding their inherent decision-making processes. Last, the evaluation of simulation outcomes has predominantly focused on

the quantitative analysis of decision types (e.g., frequency of cooperation), overlooking the LLMs' higher-level behavioral patterns that can be inferred from the temporal evolution of these decisions (Xu et al. 2024; Mao et al. 2025). The combined effect of these limitations has led to findings that are sometimes inconclusive (Brookins and DeBacker 2024; Mao et al. 2025) and contradictory (Akata et al. 2023; Fan et al. 2024), calling for more systematic evidence on the behavior of LLMs in iterated games.

In this work, we investigate the adaptability of LLMs in terms of their cooperative behavior when facing a spectrum of hostility in an iterated game scenario. We evaluate Llama2 (Touvron et al. 2023), Llama3, and GPT3.5 (Brown et al. 2020) in the Iterated Prisoner's Dilemma (Osborne and Rubinstein 1994) against adversaries with different propensities for betrayal. Our contribution is threefold. First, we introduce a meta-prompting technique designed to evaluate an LLM's comprehension of the game's rules and its ability to parse historical gameplay logs for decision-making. Second, we conduct extensive simulations over 100 rounds and determine the optimal memory span that enables the LLMs to adhere to the strategic framework of the game. Third, we analyze the behavioral patterns exhibited by the LLMs, aligning them with the core dimensions and strategies delineated in Robert Axelrod's influential research on the evolution of cooperation within strategic interactions (Axelrod and Hamilton 1981).

We observe that, overall, the three models tend to be more cooperative than humans, but they display some variability in their strategies. This variability persists even when the models are exposed to the same environment, game, and task framing. Both Llama2 and GPT3.5 displayed a more marked propensity towards cooperation than what existing literature reports about human players, indicating a favorable alignment with positive values. In contrast, Llama3 adopts a more strategic and exploitative approach that is more similar to that of humans. This approach may be advantageous in competitive scenarios where raw performance is critical, but it can be a disadvantage when it comes to aligning with positive values.

Overall, our work contributes to defining a more principled approach to using LLMs for iterated games. It makes a step towards a more systematic way to use simulations of game theoretical scenarios to probe the inherent social biases of LLMs, which might prove useful for LLM auditing and alignment (Shen et al. 2023; Mökander et al. 2023).

2 Background on Prisoner's Dilemma

2.1 Game Setup

The Prisoner's Dilemma is a classic thought experiment in Game Theory. It serves as a paradigm for analyzing conflict and cooperation between two players (Tucker and Straffin Jr 1983). In the game, the two players cannot communicate, and must independently choose between two actions: *Cooperate*, or *Defect*. Once both players have chosen their actions, payoffs are distributed based on the resulting combination of their choices. Mutual cooperation yields a reward R for each player. If one defects while the other cooperates, the defector receives a higher 'temptation' payoff T, while the cooperating player incurs a lower 'sucker's' payoff S. If both parties choose to defect, they each receive a punishment payoff P for failing to cooperate. The classical structure of the game is defined by the payoff hierarchy $T > R > P > S$, which theoretically incentivizes rational players to consistently choose defection as their dominant strategy (Axelrod 1981). In the iterated version of the game, multiple rounds are played, and the payoffs are revealed to the players at every round (Tucker and Luce 1959). The iterative nature of the game allows the players to consider past outcomes to strategically inform future actions. When humans play the game, psychological factors such as reputation and trust significantly influence the decision-making process, often leading to higher rates of cooperation than would be expected from purely rational agents (Dal Bó and Fréchette 2011; Romero and Rosokha 2018).

2.2 Strategies

In the Iterated Prisoner's Dilemma (IPD), a *strategy* refers to an algorithm that a player uses to decide their next action, taking into account the historical context of the game (Kuhn 2024). No single strategy universally outperforms all others; however, some are more effective against a broader variety of opposing strategies (Dal Bó and Fréchette 2011). This concept was demonstrated in 1980 by Robert Axelrod (Axelrod 1980), who ran an IPD tournament with multiple competing strategies. Follow-up tournaments have further diversified the range of strategies (Stewart and Plotkin 2012). Considering previous literature (Fudenberg, Rand, and Dreber 2012; Dal Bó and Fréchette 2011; Romero and Rosokha 2018), we consider the strategies that better represent the ones adopted by humans, covering more than 75% of the experimental samples in those studies. Those strategies are the following:

- Always Cooperate (AC).
- Always Defect (AD).
- Random (RND). Chooses Cooperate or Defect at random with equal probability at each round.
- Unfair Random (URNDp). Variation of Random where the probability of choosing to Cooperate is p.
- Tit For Tat (TFT). Starts with Cooperation in the first round, then mimics the opponent's previous action throughout the game.
- Suspicious Tit For Tat (STFT). A TFT strategy that begins with Defect in the first round.
- Grim Trigger (GRIM). Chooses Cooperate until the opponent defects, then chooses only Defect for the rest of the game.
- Win–Stay Lose–Shift (WSLS). Repeats the previous action if it resulted in the highest payoffs (R or T), otherwise changes action.

Tit For Tat emerged as the winning strategy in Axelrod's tournament. It is commonly observed that human players tend to favor straightforward strategies such as AD, TFT, or GRIM (Romero and Rosokha 2018). To describe the LLM

behavior in terms of these known strategies, in our experiments we calculate the similarity of the LLM's game sequences with the sequences that these hardcoded strategies would generate when playing against the same opponent.

2.3 Behavioral Dimensions

To identify the defining factors of different strategies, we combine dimensions already defined in prior studies that quantify salient behavioral properties of players based on the observed game sequences (Axelrod 1980; Mei et al. 2024):

- Nice. Propensity to not be the first to defect. For a single instance of the Iterated Prisoner's Dilemma, it is defined as 1 if the player is not the first to defect, 0 otherwise.
- Forgiving. Propensity to cooperate again after an opponent's defection, defined as: $\frac{\#\text{forgiven\_defection}}{\#\text{opponent\_defection}+\#\text{penalties}}$, where the number of 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. Propensity of defecting immediately after an opponent's uncalled defection, defined as: $\frac{\#\text{reactions}}{\#\text{provocations}}$
- Troublemaking. Propensity to defect unprovoked, defined as a counterpart of being retaliatory: $\frac{\#\text{uncalled\_defection}}{\#\text{occasions\_to\_provoke}}$, where an uncalled defection is a defection following a cooperation (or being the first action of the game) and an occasion to provoke is any cooperation from the opponent in the previous round.
- Emulative. Propensity to copy the opponent's last move: $\frac{\#\text{mimic}}{N-1}$, where a *mimic* occurs any time the player played the same action that the opponent played in the previous round and N is the number of iterations of the game.

Strategies that are Nice, Forgiving, and Retaliatory (e.g., TFT) perform best against a wide variety of opponents. Human players tend to be particularly uncooperative when faced with games where the reward R for cooperating is much lower than the temptation T to betray the other player. In the indefinitely iterated version of the Prisoner's Dilemma with a fixed probability at every round for the game to terminate, human subjects tend to be more cooperative when the probability of ending the game is low. Usually, in games with a low chance of continuation and a big gap between R and T, the majority of the strategies adopted are most similar to Always Defect (70% to 90%); in games with a longer potential time horizon and R closer to T, humans tend to be more forgiving and Tit For Tat explains a larger portion of the subjects' strategies (Dal Bó and Fréchette 2011).

3 Experimental Design

3.1 LLM Setup

In our experiments, we use `Llama-2-70b-chat-hf` and `Llama-3-70B-Instruct` as open-source language models developed by Meta and released under commercial use licenses[1,2]. We access these models through the Hugging Face platform, using their Inference API[3]. As a closed-source language model, we use `GPT-3.5-Turbo`, developed and hosted by OpenAI, accessed via their proprietary API[4]. GPT3.5 has been used in many early experiments on LLM agents in game-theoretical scenarios (Mei et al. 2024; Lorè and Heydari 2024; Xu et al. 2024).

The models are initialized with a temperature value of 0.7, which is equal to the default value for GPT models and is consistent with previous studies using models from the Llama family (Lorè and Heydari 2024; Xu et al. 2024). An analysis of robustness to different temperature settings (0.1 and 1.0) is provided in the Appendix (Figure A9).

We make the LLMs play a series of Iterated Prisoner's Dilemma games, each consisting of $N = 100$ rounds. Due to the stochastic nature of the responses that LLMs generate, we repeat each game $k = 100$ times and report the average results along with 95% confidence intervals. To evaluate the models' adaptability to different degrees of environmental hostility, we repeat the experiment by matching them against URNDp opponents (defined in §2.2) 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:

$$G_k^\alpha = [(A_i, B_i)]_{i\in[1,N]}. \tag{1}$$

From the G_k^α sequence, we extract the empirical probability of the LLM to cooperate at round i, calculated as the fraction of i^{th} rounds in which the LLM cooperated over k trials:

$$p_{coop}^\alpha(i) = \frac{1}{k}\sum_k G_k^\alpha(A_i). \tag{2}$$

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

$$p_{coop}^\alpha = \frac{1}{N}\sum_{i=1}^{N} p_{coop}^\alpha(i). \tag{3}$$

3.2 Prompting

To implement the game, we design a fixed system prompt that outlines the game's *rules*, including the payoff structure, and the player's *objective* to *'get the highest possible number of points in the long run'*. The variable part of the prompt includes the *memory* of the game, namely a log of the history of the players' actions up to the current round, along with instructions for generating the next action. The complete prompt can be found in the Appendix (Figures A1, A2, and A3). In iterated games, the information from earlier rounds is essential for a player to deduce the opponent's strategy, and adapt accordingly. Early research involving LLMs in iterated games experimented with a limited number of rounds, precluding any analysis of how the size of the memory window influences the agent's behavior (Akata et al. 2023; Guo 2023; Xu et al. 2024). In formulating the memory component, we explore various window sizes to provide the model

[1]https://ai.meta.com/llama/license/

[2]https://llama.meta.com/llama3/license/

[3]https://huggingface.co/inference-api/serverless

[4]https://openai.com/index/openai-api/

	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 used in meta-prompting to verify the LLM's comprehension of the prompt.

with information from only the n most recent rounds. We evaluate the effect of different memory window sizes by testing the LLM against an Always Defect opponent and identifying the optimal window size (see §4.2). This assessment is based on the premise that, once the LLM has gathered sufficient information to recognize the opponent's consistently defecting behavior, its actions should align with defection, which is the only logical strategy.

3.3 Meta-Prompting

The development of effective LLM prompts is an ever-evolving practice. Although certain studies suggested guidelines for prompt development (Ziems et al. 2024), achieving a consensus on the most effective prompting strategies across tasks remains challenging. Typically, the prompt quality is assessed empirically based on downstream performance (Lester, Al-Rfou, and Constant 2021). This method is suitable for conventional classification or regression tasks where some form of ground truth is clearly defined. However, it is less applicable to generative tasks that lack a formal standard of correctness. In the specific context of the Prisoner's Dilemma, any sequence of Cooperate and Defect actions could be considered plausible. This ambiguity makes 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 (Xu, Jain, and Kankanhalli 2024). Prior research involving LLMs in Game Theory experiments has assessed outputs by requesting that the LLM provide a reasoned explanation of its output (Guo 2023). However, this approach offers only a retrospective justification, which can itself suffer from hallucinations if the LLM has not fully grasped the task's underlying instructions.

To partially mitigate this issue, we introduce a novel meta-prompting technique to measure the LLMs' comprehension of a given prompt, to inform the process of prompt refinement. Specifically, we formulate a set of *prompt comprehension* questions that address three key aspects of the prompt (see Table 1): the *game rules*, to verify the LLM's grasp of the game mechanics (e.g., *'What is the lowest payoff that player A can get in a single round?'*); the chronological sequence of actions within the game history (e.g., *'Which action did player A play in round 5?'*); and the cumulative game statistics (e.g., *'What is player's B current total payoff?'*). To assess the LLMs' proficiency in responding to meta-prompting questions, we conduct a series of 3 games of 100 rounds each against RND opponents. We pose the questions at each round and compute the average accuracy of the LLMs' responses. At any given round i, a question template is instantiated into a set of questions that cover all possible combinations of the template's parameters. For example, at round i, questions referring to specific rounds (`action`$_i$ and `points`$_i$) are asked for all past rounds from 1 to $i-1$.

3.4 Behavioral Profiling

We profile the LLMs' behavior with respect to a game history G_k^α in two ways. First, we quantify behavioral patterns by computing the behavioral dimensions outlined in §2.3. This computation results in a five-dimensional numerical vector that encapsulates the behavioral characteristics of the LLMs. Second, we use the Strategy Frequency Estimation Method (SFEM) defined in previous work (Romero and Rosokha 2018) to calculate the affinity between a player's game history and any of the classic strategies used in Prisoner Dilemma tournaments (see §2.2). SFEM is a finite-mixture approach that uses likelihood maximization to estimate the likelihood of a strategy being represented in experimental data (Romero and Rosokha 2018). Given a game history, SFEM outputs a score between 0 and 1 for each strategy in the set of theoretical strategies considered. Being a mixture approach, the sum of all the SFEM scores over the set of possible strategies does not have to sum up to 1. The final SFEM scores we report are averaged over all the histories analyzed.

4 Results

4.1 LLM's Prompt Comprehension

Figure 1 presents the accuracy of responses to the prompt comprehension questions associated with the final prompt that we used in our experiments, with results averaged over all trial runs. Overall, the models exhibit a good understanding of the concepts assessed by the questions, with most responses achieving an accuracy ranging from 0.8 to 1.0.

We iteratively tested multiple prompt versions against the response accuracy of Llama2. No further iterations of other models were needed, as the best prompt for Llama2 yielded very high response accuracy in both Llama3 and GPT3.5. Generally, adding explicit information about the game state and rules led to a better level of prompt comprehension. For

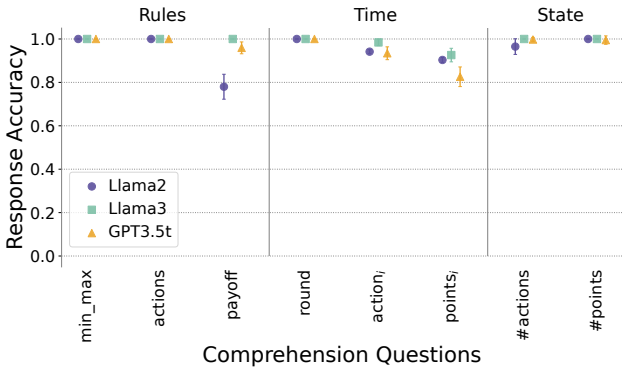

Figure 1: Accuracy of the models' responses to the prompt comprehension questions defined in Table 1. The questions are categorized into three groups, each assessing different aspects: the rules of the game, its temporal evolution, and its current state. We show 95% confidence intervals, computed from 100 games.

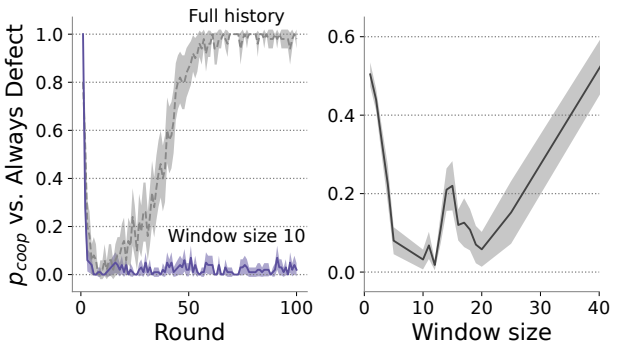

Figure 2: *Left*: Llama2's probability of cooperation (p_{coop}) against an Always Defect opponent, when using a memory window size of 10 vs. including the full game history in the prompt. *Right*: steady-state probability calculated on the last 10 rounds of p_{coop} for different memory windows sizes. We show 95% confidence intervals, computed from 100 games.

illustration, Figure A7 in the Appendix includes a comparative analysis of accuracy scores derived from an earlier version of the prompt, which lacked a summary of the cumulative point totals for the players. In the absence of explicit total counts, Llama2 was required to sum all points from the game history to determine the total tally, which significantly impacted its performance on the `#actions` and `#points` questions. This limitation aligns with previous findings that highlight that LLMs tend to struggle with arithmetic (Xu et al. 2024; Aher, Arriaga, and Kalai 2023; Wei et al. 2022). The explicit inclusion of the sum of scores into the prompt markedly improved performance, achieving near-perfect accuracy.

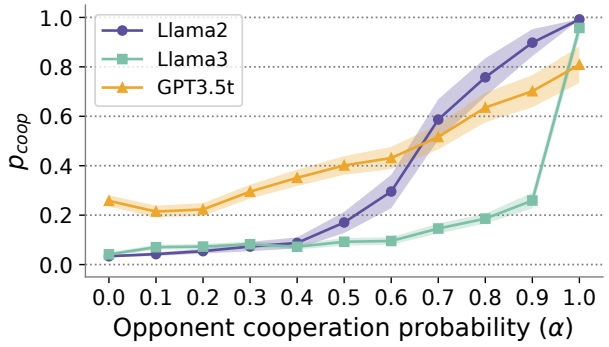

Figure 3: Models' probability of cooperation (p_{coop}) against Unfair Random opponents with increasing cooperation probability α. We show 95% confidence intervals, computed from 100 games.

4.2 Effect of Memory Window Size

Figure 2 (left) shows the probability of the LLM cooperating at each round in games of 100 rounds, considering two memory window sizes: $m = 100$ and $m = 10$. Under both conditions, the LLM shifts towards a stance of consistent defection after approximately 5 to 10 rounds. Notably, this duration equals or exceeds the maximum number of rounds considered in previous studies. Without any constraints on the history length, the LLM's cooperation level quickly starts rising back, converging to full cooperation after the 50^{th} round. This pattern may be attributed to a combination of two factors: Llama2's intrinsic preference for positive constructs (Lorè and Heydari 2024) (favoring cooperation over defection) and its limited effectiveness in extracting actionable insights from long prompts (Xi et al. 2025). With a memory window restricted to the 10 most recent rounds, the LLM retains a full-defection stance throughout the game. We replicated this experiment across a range of memory window sizes and determined their respective equilibrium states by calculating the average cooperation probability in the final 10 rounds (from the 90^{th} to the 100^{th}). As illustrated in Figure 2 (right), memory windows size around 10 yield the expected outcome. Therefore, we select a window size of $m = 10$ for the remainder of our experiments. Differently, Llama3 and GPT3.5 show no variation in their cooperation levels regardless of the memory window size (see Figure A8 in Appendix), allowing us to use the same prompt designed for Llama2. Notably, Llama3 does not suffer from the same bias towards cooperation as Llama2, and keeps its defective behavior also when provided with the full game history. GPT3.5 exhibits the less optimal behavior, converging towards defection much more slowly.

4.3 Behavioral Patterns

Probability of Cooperation We examine the overall propensity for cooperation exhibited by the LLM across various degrees of environmental hostility. Figure 3 shows the relationship between the probability of cooperative behavior p^{α}_{coop} of each model and the varying cooperation levels α of

an Unfair Random opponent.

Among the three models, Llama3 demonstrates the most strategic approach. It maintains a very low level of cooperation even when the opponent is nearly always cooperating ($p_{coop} < 0.3$ for every $\alpha < 1$), but it increases its cooperation to nearly 100% when the opponent is AC ($\alpha = 1$). In contrast, Llama2's behavior follows a sigmoidal curve across the entire range of α from 0 to 1. This indicates a rapid transition from a predominantly defecting strategy to a more cooperative attitude. The sigmoid curve is characterized by a relatively flat left tail, maintaining a stable probability of cooperation near $p_{coop} = 0$ for α values up to 0.4. The curve's inflection point is between 0.6 and 0.7—well beyond 0.5—suggesting a cautious approach in interpreting the opponent's actions. Compared to Llama3, the older Meta model is less guarded and more prone to increase its cooperation as soon as the opponent's cooperation significantly rises over the 50/50 chance. GPT3.5 displays a less 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.

In general, the models exhibit non-linear behaviors but with different characteristics. A common trend is the increasing cooperation as α grows, demonstrating a minimum level of strategic behavior for all LLMs.

These findings, especially concerning Llama3, support previous research that highlights the cautious response patterns of LLMs in repeated game scenarios (Akata et al. 2023; Phelps and Russell 2023).

SFEM Profile The probability of cooperation provides a macroscopic perspective on a player's actions. However, to capture more nuanced strategic patterns that emerge during the match, we employed SFEM analysis (defined in §3.4) to estimate the similarity between the behavioral patterns exhibited by the LLMs and those commonly observed in games involving human players.

Figure 4 illustrates which strategies best explain the behaviors of each model as the values of α increase. When the adversary's probability of cooperation exceeds the 0.6 threshold, there is a noticeable shift in both Llama2's and GPT3.5's strategy from Grim Trigger to Always Cooperate. In contrast, Llama3 consistently aligns with the Grim Trigger strategy across all values of α, displaying an exploitative approach even when the opponent tends to cooperate for the majority of the time. SFEM analysis further indicates that these strategies are the most representative of the LLMs' strategic behavior, with no other strategies being significantly indicative.

Previous studies have shown that humans tend to be particularly uncooperative when the reward R is much lower than the temptation T and when playing for shorter time horizons, typically opting for startegies close to Always Defect (Dal Bó and Fréchette 2011; Romero and Rosokha 2018). Conversely, humans become more cooperative when the reward is much higher and the time horizon is longer, opting for a mixture of Tit For Tat and Grim Trigger strategies.

Comparing this prior knowledge to our results indicates

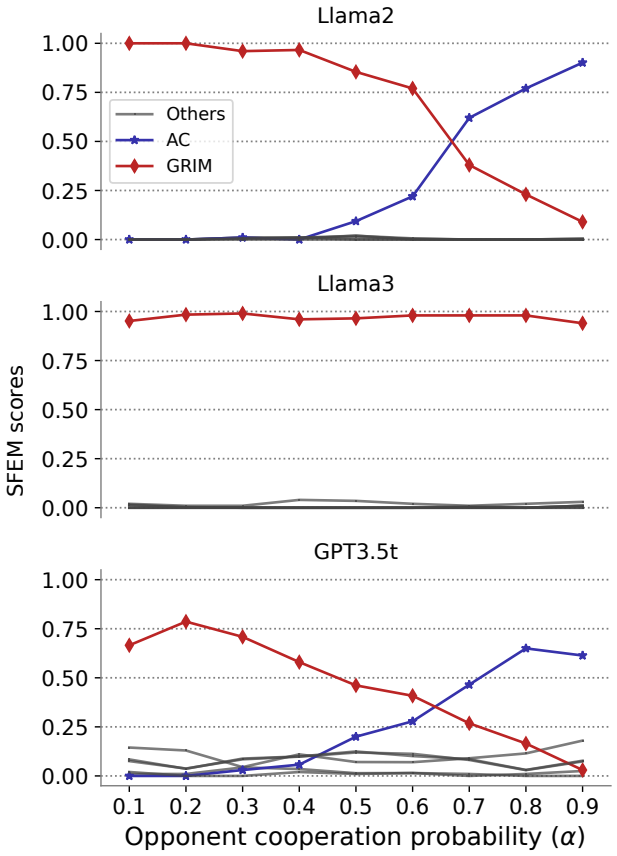

Figure 4: SFEM scores quantifying the similarity between the models' sequences of actions and known strategies adopted in the Iterated Prisoner's Dilemma game (defined in §2.2). The models' behavioral sequences come from games against Unfair Random opponents with increasing cooperation probability α. Some SFEM scores are not shown because they are not well-defined for extreme values of α.

that GPT3.5 and Llama2 are consistently more cooperative than humans. In situations that do not favor cooperation—where the opponent defects more frequently than cooperates—they adopt the GRIM strategy, unlike humans who tend to use the AD strategy. Similarly, when the context supports cooperation (i.e., the opponent cooperates more than half of the time), both GPT3.5 and Llama2 employ the AC strategy, which is notably more cooperative than the human preference for TFT or GRIM. On the other hand, Llama3 is more cooperative than humans in hostile environments ($p_{\text{coop}} < 0.5$), also using the GRIM strategy. However, in environments that encourage cooperation, Llama3 behaves more like humans by continuing to use the GRIM strategy, whereas humans typically mix between the more forgiving TFT strategy and GRIM.

Behavioral Profile At a finer level of analysis, we characterize the behavior of the LLMs along the dimensions outlined in Section 2.3 (Figure 5).

When the parameter α is set to low values, the models exhibit highly uncooperative traits: they frequently retaliate following instances of betrayal, seldom revert to cooperative behavior after defecting, and are often the initiators of unprovoked defections. GPT3.5 exhibits these traits in a more moderate manner, whereas Llama2 and Llama3 display more extreme profiles in this regard. More noticeable differences among the models' responses emerge beyond $\alpha = 0.5$. While GPT3.5 and Llama2 generate choices that are more Forgiving and less Troublemaking, Llama3 remains Troublemaking and rarely Forgiving for every $\alpha < 1.0$. On the other hand, the three models are constantly Nice across conditions, rarely defecting first. GPT3.5 shows the lowest scores but remains above 0.5 even against an opponent that never defects ($\alpha = 1$). In this scenario, all the LLMs never defect throughout the whole game in 50% of the runs. Interestingly, there is no scenario in which any LLM is simultaneously Nice, Forgiving, and Retaliatory—the three conditions that characterize the most successful strategies, such as Tit For Tat.

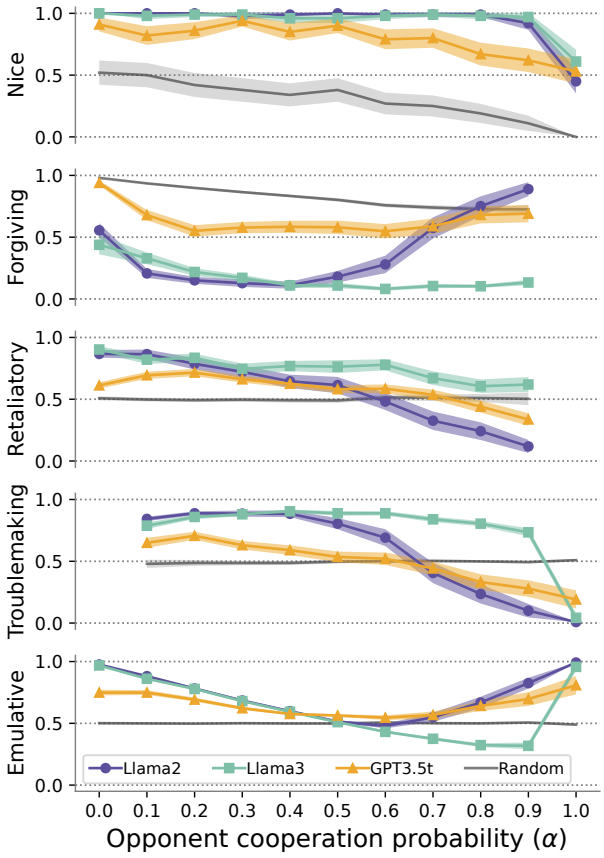

Figure 5: Presence of behavioral traits in the models' actions when playing against Unfair Random opponents with increasing cooperation probability α. The values of the same traits calculated for a Random agent playing against Unfair Random opponents are reported. Some traits are not defined for extreme values of α. We show 95% confidence intervals, computed from 100 games.

5 Discussion and Conclusion

Our study contributes to the broad literature on behavioral studies of Large Language Models as artificial social agents. Specifically, we study the responses of Llama2, Llama3, and GPT3.5 to the prototypical social scenario of the Prisoner's Dilemma. Building upon prior studies that explored the application of LLMs in classic Game Theory experiments, our work introduces a more systematic experimental setup that incorporates quantitative checks to better align the LLMs' responses to the complex task description. We have shown that aspects like prompt comprehension, memory representation, and duration of the simulation play crucial roles, especially for less powerful models, with the potential to significantly distort the experimental outcomes, if left unchecked. Our framework provides a quantitative guide for selecting the simulation variables and improving the prompt.

Our findings contribute a new benchmark to the literature exploring the outcomes of iterated games, both among humans and among AI agents. Moreover, our work adds a new perspective to the emergent body of work studying population biases of social AI agents (Coppolillo, Manco, and Aiello 2025; Flint Ashery et al. 2025). In contrast to the behavioral patterns observed in humans playing the Prisoner's Dilemma game (Dal Bó and Fréchette 2011), the three models displayed a more marked overall propensity towards cooperation. Under conditions that disincentivize cooperation, most human players adopt a stance of complete defection, whereas the LLMs' strategy, albeit mostly uncooperative, is characterized by an initial trust in the opponent's cooperation (Nice), reminiscent of the strategy known in Game Theory as Grim Trigger. When the environment is more favorable to cooperative play, Llama3 playing Grim Trigger becomes comparable to human strategies that often resemble either Grim Trigger or Tit For Tat. Differently, Llama2 and GPT3.5 tend towards a consistently cooperative approach. Notably, Llama2's shift from Grim Trigger to Always Cooperate occurs quite abruptly as the opponent's defection probability drops below 30%, while GPT3.5's transition is smoother and starts when α exceeds 0.2.

Our results are in line with early experiments involving LLMs, which indicated a tendency for these models to cooperate in repeated games (Brookins and DeBacker 2024; Mei et al. 2024). However, the broader research in this area has yielded mixed outcomes (Akata et al. 2023; Phelps and Russell 2023). Distinct from previous studies, our findings are derived from extensive game simulations conducted over numerous rounds and benefit from an experimental framework that leverages quantitative checks for accuracy.

Overall, our findings offer a robust baseline for understanding LLMs' behaviors in the Iterated Prisoner's Dilemma (IPD), a widely used Game Theory experiment for assessing agents' responses to conflictual scenarios. Establishing this baseline allows for clearer differentiation between the inherent tendencies of the models and the effects of specific elements within the game setting.

However, it is important to acknowledge the limitations of our work, which opens the way for further research. First, our analysis was conducted using three models, specifically

Llama2, Llama3, and GPT3.5, which, at the time of writing, are among the most advanced models available (Touvron et al. 2023; Brown et al. 2020)[5]. Nevertheless, the field is progressing at an unprecedented pace, with new models being introduced regularly. To determine whether the behavioral patterns observed in our study are consistent across many different models, it is important to conduct comparative analyses with models that vary significantly in terms of parameter size and the volume of their training data. Second, our study's scope was limited to assessing the LLM's responses to *random* strategies, and with a fixed payoff structure. Exploring the LLM's interactions with more sophisticated opponents would enable us to better delineate the boundaries of LLMs' inferential abilities in social contexts, and to draw more detailed behavioral profiles under a broader spectrum of conditions that more comprehensively represent prototypical social scenarios. We explored only the use of Zero-Shot Chain-of-Thought techniques without obtaining any improvement. Other options like Tree-of-Thought (Yao et al. 2023) techniques or external modules (Wang et al. 2024) can be employed to test different reasoning conditions for the models. Furthermore, the experimental framework of our study considers only a single LLM agent. Creating social groups of AI agents that interact through iterated games like The Prisoner's Dilemma would open up a wealth of opportunities to study emergent behaviors in synthetic societies, an avenue of research that is increasingly recognized as fundamental for a proper understanding of how LLMs can affect human societies (Bail 2024). Last, despite the numerical guidelines we implemented to evaluate the quality of the prompt, our refinements of the prompt were not guided by any principle other than experience and instinct. Our attempt to use Chain-of-Thought (Kojima et al. 2022; Zhou et al. 2023) as a structured way to approach prompt revision resulted in prompts that did not improve performance in our prompt comprehension question inventory (see Figure A6 in Appendix). In this respect, our work provides yet another example of how prompt engineering would benefit from supporting tools to constrain its highly discretionary nature. Another interesting direction for future research is to explore persona prompting (Hu and Collier 2024), where models are guided towards specific behaviors, such as altruism or selfishness. However, in this study, we intentionally avoided using personas to influence the outputs of the LLMs, as our primary objective is to assess the models' *inherent biases* towards different behavioral patterns in game-theoretical scenarios. Similarly, models could be enhanced by incorporating a planning module that considers the potential outcomes of various actions over multiple steps (Kambhampati et al. 2024).

Despite its limitations, our work has two main implications. From the theoretical perspective, it expands our knowledge of the inherent biases of LLMs in social situations, which is crucial to inform their deployment across different contexts. From the practical perspective, it provides a principled way to approach game theoretical simulations with LLMs. This constitutes a step towards using these simulations as reliable and reproducible tools that could be used as tests to verify LLM alignment to desired principles of social cooperation (Shen et al. 2023).

[5]https://ai.meta.com/blog/meta-llama-3/

6 Related Work

Next, we briefly review previous work using LLMs for social reasoning, the generation of human-like synthetic data, and simulations of human behavior.

6.1 LLMs as Agents

Argyle et al. (2023) instructed LLMs to answer surveys as if they belonged to specified socio-demographic groups. They showed a high similarity between the responses generated by the LLM and those provided by the demographic groups it was asked to emulate. LLMs impersonating human agents with different profiles were used to explore the negotiation abilities of the models (Davidson et al. 2024) or to create synthetic social networks, to observe emergent social behavior, most notably opinion dynamics and information spreading (Chuang et al. 2024; Marzo, Pietronero, and Garcia 2023; He et al. 2024). Park et al. (2023) developed a society with synthetic agents interacting with elements of a synthetic world, showed that those agents were able to adopt behaviors that are typical of humans without being directly prompted to do so. Using the same framework, Ren et al. (2024) showed that those agents were also able to build and spread social norms, while Piatti et al. (2024) developed a similar framework to investigate the robustness of LLMs societies.

6.2 LLMs in Game Theory

Early work on the application of LLMs to Game Theory experiments touched upon both 1-time and iterated games. Single-iteration experiments offer limited insight into LLM behavior. Brookins and DeBacker (2024) showed that LLMs are more biased towards fairness and cooperation when compared to a human baseline sample. In contrast, Aher, Arriaga, and Kalai (2023) found that there is an overall alignment between the LLM-based agent behavior and the human ones. Investigating the ability of LLMs to predict human choices in 1-time games, Capraro, Paolo, and Pizziol (2024) showed that only more powerful models are capable of doing it, although they overestimate the altruistic tendency of human players. When focusing on iterated games, the spectrum of patterns that can be identified expands, allowing more refined analysis. For example, Akata et al. (2023) managed to identify that LLMs can be particularly unforgiving. Mei et al. (2024) discovered instead that the same models show a higher cooperation rate than compared to humans. Fan et al. (2024) exploited the iteration of games to check the level of the opponent's strategy that the LLM was able to infer from the history of actions. They showed that the inference capability of the LLM is limited, calling for more systematic approaches to structure memory and prompts. Testing a novel benchmark with different games, Duan et al. (2024) studied multiple LLMs showing that closed-source models tend to achieve better performance than open-source ones.

Although they did not examine the behavioral characteristics of the LLMs' responses, they also found that Zero-Shot Chain-of-Thought techniques are not always beneficial, which aligns with our results.

Code and Data Availability

All code and game traces are available on GitHub at https://github.com/NicoloFontana/nicer_than_humans_icwsm25.

Acknowledgments

This work was partially supported by the Italian Ministry of Education (PRIN grant DEMON prot. 2022BAXSPY) and the European Union (NextGenerationEU project PNRR-PE-AI FAIR). NF acknowledges 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". LMA acknowledges the support from the Carlsberg Foundation through the COCOONS project (CF21-0432).

References

Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning*, 337–371. PMLR.

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.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3): 337–351.

Axelrod, R. 1980. Effective Choice in the Prisoner's Dilemma. *Journal of Conflict Resolution*, 24(1): 3–25.

Axelrod, R. 1981. The Emergence of Cooperation among Egoists. *American Political Science Review*, 75(2): 306–318.

Axelrod, R.; and Hamilton, W. D. 1981. The evolution of cooperation. *science*, 211(4489): 1390–1396.

Bail, C. A. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21): e2314021121.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Breum, S. M.; Egdal, D. V.; Mortensen, V. G.; Møller, A. G.; and Aiello, L. M. 2024. The Persuasive Power of Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18: 152–163.

Brookins, P.; and DeBacker, J. 2024. Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Economics Bulletin*, 44(1): 25–37.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Camerer, C. F. 1997. Progress in Behavioral Game Theory. *Journal of Economic Perspectives*, 11(4): 167–188.

Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P.; and Sun, L. 2025. A Survey of AI-Generated Content (AIGC). *ACM Comput. Surv.*, 57(5): 125:1–125:38.

Capraro, V.; Paolo, R. D.; and Pizziol, V. 2024. Assessing Large Language Models' ability to predict how humans balance self-interest and the interest of others. arXiv:2307.12776.

Chuang, Y.-S.; Goyal, A.; Harlalka, N.; Suresh, S.; Hawkins, R.; Yang, S.; Shah, D.; Hu, J.; and Rogers, T. 2024. Simulating Opinion Dynamics with Networks of LLM-based Agents. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 3326–3346. Mexico City, Mexico: Association for Computational Linguistics.

Coppolillo, E.; Manco, G.; and Aiello, L. M. 2025. Unmasking Conversational Bias in AI Multiagent Systems. *arXiv preprint arXiv:2501.14844*.

Dafoe, A.; Hughes, E.; Bachrach, Y.; Collins, T.; McKee, K. R.; Leibo, J. Z.; Larson, K.; and Graepel, T. 2020. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*.

Dal Bó, P.; and Fréchette, G. R. 2011. The Evolution of Cooperation in Infinitely Repeated Games: Experimental Evidence. *American Economic Review*, 101(1): 411–429.

Davidson, T. R.; Veselovsky, V.; Kosinski, M.; and West, R. 2024. Evaluating Language Model Agency Through Negotiations. *The Twelfth International Conference on Learning Representations*.

Duan, J.; Zhang, R.; Diffenderfer, J.; Kailkhura, B.; Sun, L.; Stengel-Eskin, E.; Bansal, M.; Chen, T.; and Xu, K. 2024. GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations. arXiv:2402.12348.

Fan, C.; Chen, J.; Jin, Y.; and He, H. 2024. Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1616): 17960–17967.

Ferrara, E. 2024. GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, 1–21.

Flint Ashery, A.; Aiello, L. M.; ; and Baronchelli, A. 2025. Emergent Social Conventions and Collective Bias in LLM Populations. *Science Advances*.

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694.

Fontana, N.; Corso, F.; Zuccolotto, E.; and Pierri, F. 2025. Evaluating open-source Large Language Models for automated fact-checking. arXiv:2503.05565.

Fudenberg, D.; Rand, D. G.; and Dreber, A. 2012. Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World. *American Economic Review*, 102(2): 720–749.

Guo, F. 2023. GPT in Game Theory Experiments. arXiv:2305.05516.

He, J. K.; Wallis, F. P. S.; Gvirtz, A.; and Rathje, S. 2024. Artificial intelligence chatbots mimic human collective behaviour. *British journal of psychology (London, England*.

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An Overview of Catastrophic AI Risks. *arXiv preprint arXiv:2306.12001*.

Horton, J. J. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? DOI: 10.3386/w31122.

Hu, T.; and Collier, N. 2024. Quantifying the Persona Effect in LLM Simulations. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10289–10307. Bangkok, Thailand: Association for Computational Linguistics.

Kambhampati, S.; Valmeekam, K.; Guan, L.; Verma, M.; Stechly, K.; Bhambri, S.; Saldyt, L. P.; and Murthy, A. B. 2024. Position: LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. In *Proceedings of the 41st International Conference on Machine Learning*, 22895–22907. PMLR.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35: 22199–22213.

Kuhn, S. 2024. Prisoner's Dilemma: Strategies for the Iterated Prisoner's Dilemma. https://plato.stanford.edu/archives/win2024/entries/prisoner-dilemma/. Accessed: 2025-04-09.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Liu, G.; Bono, C. A.; and Pierri, F. 2025. Comparing diversity, negativity, and stereotypes in Chinese-language AI technologies: an investigation of Baidu, Ernie and Qwen. *PeerJ Computer Science*, 11: e2694.

Lorè, N.; and Heydari, B. 2024. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1): 18490.

Mao, S.; Cai, Y.; Xia, Y.; Wu, W.; Wang, X.; Wang, F.; Guan, Q.; Ge, T.; and Wei, F. 2025. ALYMPICS: LLM Agents Meet Game Theory. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 2845–2866. Abu Dhabi, UAE: Association for Computational Linguistics.

Marzo, G. D.; Pietronero, L.; and Garcia, D. 2023. Emergence of Scale-Free Networks in Social Interactions among Large Language Models. arXiv:2312.06619.

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.

Mökander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2023. Auditing large language models: a three-layered approach. *AI and Ethics*, 1–31.

Nogara, G.; Pierri, F.; Cresci, S.; Luceri, L.; Törnberg, P.; and Giordano, S. 2025. Toxic Bias: Perspective API Misreads German as More Toxic. *Proceedings of the International AAAI Conference on Web and Social Media*.

Osborne, M. J.; and Rubinstein, A. 1994. *A course in game theory*. MIT press.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, 1–22. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701320.

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

Piatti, G.; Jin, Z.; Kleiman-Weiner, M.; Schölkopf, B.; Sachan, M.; and Mihalcea, R. 2024. Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. *Advances in Neural Information Processing Systems*, 37: 111715–111759.

Ren, S.; Cui, Z.; Song, R.; Wang, Z.; and Hu, S. 2024. Emergence of Social Norms in Generative Agent Societies: Principles and Architecture. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence Human-Centred AI*, volume 8, 7895–7903.

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

Saffari, H.; Shafiei, M.; Rooein, D.; Pierri, F.; and Nozza, D. 2025. Can I introduce my boyfriend to my grandmother? Evaluating Large Language Models Capabilities on Iranian Social Norm Classification. arXiv:2406.09123.

Schramowski, P.; Turan, C.; Andersen, N.; Rothkopf, C. A.; and Kersting, K. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3): 258–268.

Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Liu, Y.; and Xiong, D. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Stewart, A. J.; and Plotkin, J. B. 2012. Extortion and cooperation in the Prisoner's Dilemma. *Proceedings of the National Academy of Sciences*, 109(26): 10134–10135.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tucker, A. W.; and Luce, R. D. 1959. *Contributions to the Theory of Games*. Princeton University Press. ISBN 978-0-691-07937-0. Google-Books-ID: 9lSVFzsTGWsC.

Tucker, A. W.; and Straffin Jr, P. D. 1983. The Mathematics of Tucker: A Sampler. *The Two-Year College Mathematics Journal*.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; Zhao, W. X.; Wei, Z.; and Wen, J. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; Zheng, R.; Fan, X.; Wang, X.; Xiong, L.; Zhou, Y.; Wang, W.; Jiang, C.; Zou, Y.; Liu, X.; Yin, Z.; Dou, S.; Weng, R.; Qin, W.; Zheng, Y.; Qiu, X.; Huang, X.; Zhang, Q.; and Gui, T. 2025. The rise and potential of large language model based agents: a survey. *Science China Information Sciences*, 68(2): 121101.

Xu, L.; Hu, Z.; Zhou, D.; Ren, H.; Dong, Z.; Keutzer, K.; Ng, S.-K.; and Feng, J. 2024. MAgIC: Investigation of Large Language Model Powered Multi-Agent in Cognition, Adaptability, Rationality and Collaboration. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7315–7332. Miami, Florida, USA: Association for Computational Linguistics.

Xu, Z.; Jain, S.; and Kankanhalli, M. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Yang, K.-C.; and Menczer, F. 2024. Anatomy of an AI-powered malicious social botnet. *Journal of Quantitative Description: Digital Media*, 4.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in Neural Information Processing Systems*, 36: 11809–11822.

Zhou, Y.; Muresanu, A. I.; Han, Z.; Paster, K.; Pitis, S.; Chan, H.; and Ba, J. 2023. Large Language Models are Human-Level Prompt Engineers. In *The Eleventh International Conference on Learning Representations*.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1): 237–291.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, in "Experimental Design".
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we discuss biases and limitations in "Discussion and Conclusion"
   (e) Did you describe the limitations of your work? Yes, limitation are presented and discussed in "Discussion and Conclusion"
   (f) Did you discuss any potential negative societal impacts of your work? We discuss negative societal impact in "Ethical Impact".
   (g) Did you discuss any potential misuse of your work? Yes, we discuss potential misuse in "Ethical Impact".
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we share model parameters in "Experimental Design" and the full prompt in the Appendix
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA.
   (b) Did you include complete proofs of all theoretical results? NA.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. The prompts are shared in the Appendix and the code and data are available at https://github.com/NicoloFontana/nicer_than_humans_icwsm25
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, we report 95% confidence intervals.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, we use models hosted on the Hugging Face API.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, specified in "Experimental Design"
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? NA
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA.
   (d) Did you discuss how data is stored, shared, and deidentified? NA.

Ethical Impact

The deployment of Large Language Models as AI social agents raises numerous ethical considerations that are currently the subject of intense scrutiny by the interdisciplinary research community. The extraordinary capabilities of these models to generate text have led several scientists to envision alarming scenarios in which the seamless integration of AI agents into the online social discourse may facilitate the dissemination of harmful content, the spread of misinformation, and the propagation of 'semantic garbage', ultimately damaging our societies (Floridi and Chiriatti 2020; Weidinger et al. 2022; Hendrycks, Mazeika, and Woodside 2023). As a result, any research exploring the characteristics of LLMs as social agents could, directly or indirectly, contribute knowledge that might be exploited to implement and deploy LLM-based technologies for malicious purposes. While recognizing this risk, we also believe that conducting research on LLM-based agents is essential to assess potential risks and to guide efforts aimed at developing strategies to mitigate them. Our study contributes positively to deepen our understanding of how LLMs react to social stimuli.

Even when deploying LLM-based agents for ethical purposes, trade-offs between the obtained benefit and the high level of power consumption required to run them should be carefully considered (Bender et al. 2021).

Appendix

Prompts and Their Variations

Figures A1, A2 and A3 show examples of the three parts of the final prompt we used in our experiments. Figures A4 and A5 respectively show the variation in the instructing prompt when using the Zero-Shot Chain-of-Thought techniques from Kojima et al. (2022) and Zhou et al. (2023). Figure A6 displays the accuracy obtained by Llama2 in the comprehension questions when using the two variations and the unvaried prompt.

Figure A7 shows the comparison in the accuracy of Llama2 between the initial prompt without explicit score counts and the final version of it.

Effect of Memory Window Size

Figure A8 shows the probability of cooperation for Llama3 and GPT3.5 when using memory windows of different sizes. Different from Llama2, neither of these models significantly changes its behavior depending on the window size. In particular, the trend observed for GPT3.5, which tends to stabilize around the value 0.2, is in accordance with the average p_{coop} shown in Figure 5 for $\alpha = 0$

Effect of Temperature

We explore the impact of varying the temperature hyperparameter on the probability of cooperation p_{coop} of Llama3

System prompt

```
<s>[INST]
<<SYS>>
Context: Player A and player B are playing a
multi-round game.
At each turn player A and player B
simultaneously perform one of the following
actions: {"Cooperate", "Defect"}
The payoffs for each combination of chosen
actions are the following:
If A plays "Defect" and B plays "Defect", A
collects 1 points and B collects 1 points.
If A plays "Defect" and B plays "Cooperate",
A collects 5 points and B collects 0 points.
If A plays "Cooperate" and B plays "Defect",
A collects 0 points and B collects 5 points.
If A plays "Cooperate" and B plays
"Cooperate", A collects 3 points and B
collects 3 points.
They will play a total of 100 rounds of this
game.
Remember that a player's objective is to get
the highest possible number of points in the
long run.
<<SYS>>
```

Figure A1: *System prompt* describing the players' roles, the rules of the games, the number of rounds to be played, and the player's objective. This part of the prompt is fixed in all simulations.

Contextual prompt

```
The history of the game in the last 5 rounds
is the following:
Round 2: A played "Cooperate" and B
played "Defect" A collected 0 points and
B collected 5 points.
Round 3: A played "Defect" and B played
"Defect" A collected 1 points and B
collected 1 points.
Round 4: A played "Cooperate" and B
played "Defect" A collected 0 points and
B collected 5 points.
Round 5: A played "Defect" and B played
"Cooperate" A collected 5 points and B
collected 0 points.
Round 6: A played "Defect" and B played
"Defect" A collected 1 points and B
collected 1 points.
In total, A chose "Cooperate" 2 times and
chose "Defect" 3 times, B chose "Cooperate"
1 times and chose "Defect" 4 times.
In total, A collected 7 points and B
collected 12 points.
Current round: 7.
```

Figure A2: *Contextual prompt* containing information about: the game history in the last n rounds (5 in this example), the overall amount of times each player chose each action, the overall amount of points collected by each player, and the current round. For each turn, the prompt contains: the action played by each player and the points collected by each player. This prompt changes at each round.

Instructing prompt

```
Remember to use only the following JSON
format:
{"action": <ACTION_of_A>, "reason":
<YOUR_REASON>}
Answer saying which action player A should
play.
Remember to answer using the right
format.[/INST]
```

Figure A3: *Instructing prompt*. The LLM is instructed on the nature and format of the answer. This part of the prompt component is replaced with prompt comprehension questions in the phase of prompt tuning.

Instructing prompt

```
Remember to use only the following JSON
format:
{"action": <ACTION_of_A>, "reason":
<YOUR_REASON>}
Answer saying which action player A should
play.
Remember to answer using the right
format.[/INST]
Let's think step by step
```

Figure A4: Variation of the *instructing prompt* with Kojima et al. (2022) Zero-Shot Chain-of-Thought.

Instructing prompt

```
Remember to use only the following JSON
format:
{"action": <ACTION_of_A>, "reason":
<YOUR_REASON>}
Answer saying which action player A should
play.
Remember to answer using the right
format.[/INST]
Let's work this out in a step-by-step way to
be sure we have the right answer
```

Figure A5: Variation of the *instructing prompt* with Zhou et al. (2023) Zero-Shot Chain-of-Thought.

and GPT3.5[1]. For each model and each temperature (0.1 and 1.0), we run 10 games of 100 rounds each against opponents with varying cooperation probability α. Figure A9 presents the average probability of cooperation for Llama3 and GPT3.5 for three different temperature values up to 1.0. At temperatures greater than 1.0, the models tend to produce seemingly random tokens, which renders their output unusable.

The Pearson correlation between the values of p_{coop} across different temperature pairs is in the range $[0.97 - 1]$,

[1]HuggingFace discontinued access to Llama2 through their Inference API in May 2024. Since then, the model has been available only on dedicated server hosting at considerable costs. We, therefore, limit our tests on temperature to Llama3 and GPT3.5 only.

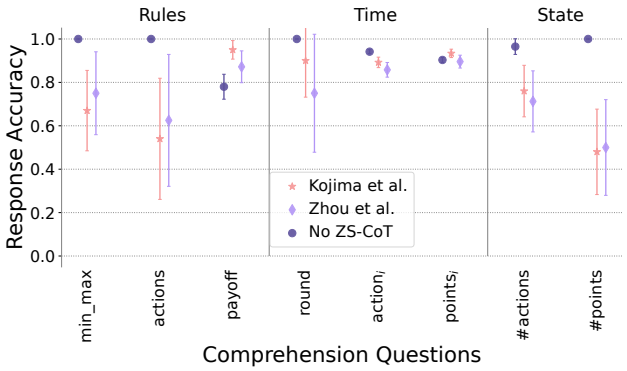

Figure A6: Accuracy of Llama2's responses to the prompt comprehension questions defined in Table 1 using two Zero-Shot Chain-of-Thought variations of the prompt compared to the accuracy obtained with the unvaried prompt. We show 95% confidence intervals, computed from 100 games.

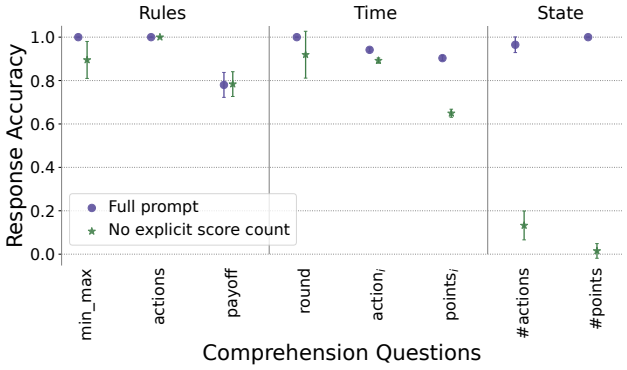

Figure A7: Accuracy of Llama2 responses to the prompt comprehension questions defined in Table 1. The accuracy of the final full prompt is compared with an earlier version of the prompt that lacked a summary of the cumulative rewards achieved by the players. We show 95% confidence intervals, computed from 100 games.

indicating that the temperature does not affect the general trends of p_{coop} as α varies. However, the growth of the cooperation curve for GPT3.5 turns from roughly linear to a sigmoid as temperature decreases. This suggests that, in some models, different levels of determinism can influence the boundaries of their decision states. In this specific case, it appears that the noise that the increased temperature adds to the generated choices smoothens the sharp transition between the uncooperative and cooperative states that can be observed at low temperatures.

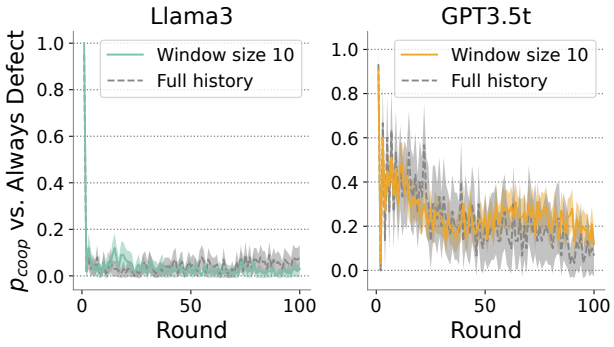

Figure A8: Llama3's and GPT3.5's probability of cooperation (p_{coop}) against an Always Defect opponent, when using a memory window size of 10 vs. including the full game history in the prompt.

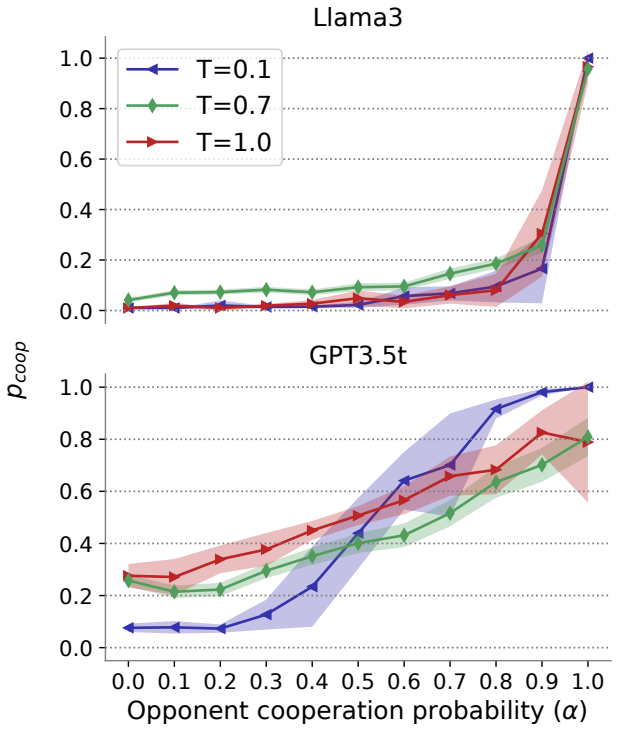

Figure A9: Llama3's and GPT3.5's probability of cooperation (p_{coop}) against Unfair Random opponents with increasing cooperation probability α for different values of the temperature hyperparameter. We show 95% confidence intervals, computed from 10 games.