

Evaluating Social Credibility of Generative Agents: Incentives, Beliefs, and Coordination

Stéphane Bonnevey

Université Lyon 1, Univ Lumière Lyon 2, ERIC, EA 3083
Villeurbanne, France
Stephane.Bonnevey@univ-lyon1.fr

Maxime Morge

Université Lyon 1, INSA Lyon, CNRS, LIRIS, UMR 5205
Villeurbanne, France
Maxime.Morge@univ-lyon1.fr

ABSTRACT

Recent advances in Large Language Models (LLMs) have enabled the development of Generative Agents (GAs) capable of decision-making and interaction. This paper investigates whether such agents can exhibit socially credible behavior. We evaluate 5 LLMs across 3 canonical game-theoretic environments. Our results show that only Mistral-Small demonstrates the closest human-like adaptation to varying incentives. While some GAs are able to anticipate their opponent’s behavior, few succeed in integrating these predictions into coherent strategic decisions. These behavioral shortcomings explain why coordination remains particularly challenging: most GAs struggle to align with others, even when communication is allowed.

KEYWORDS

Evaluation of generative agents, Cooperation and coordination

ACM Reference Format:

Stéphane Bonnevey and Maxime Morge. 2026. Evaluating Social Credibility of Generative Agents: Incentives, Beliefs, and Coordination. In *Proc. of the Strategic Engineering Workshop on LLMs and Game Theory (SE@AAMAS2026)*, Paphos, Cyprus, <https://sites.google.com/view/se-aamas2026>, May 25, 2026, IFAAMAS, 10 pages.

1 INTRODUCTION

Generative agents (GAs), based on Large Language Models (LLMs), have recently demonstrated remarkable capabilities in simulating human behavior [3, 15, 20, 23, 24, 36, 50]. These advances have sparked growing interest in their application to computational social science [17], where they are used to model, analyze, and predict social dynamics through multi-agent modeling [16, 21, 32, 33, 42–44]. GAs can adopt heterogeneous roles and objectives while engaging in collective coordination, acting as learners without prior experience yet endowed with implicit knowledge. However, one question remains open [1, 5, 6, 8, 9, 11, 12, 14, 19, 25–29, 34–37, 40, 49, 53]: is their social behavior plausible?

This paper explores the potential of *homo silicus* for social simulation by examining the behavior of intelligent machines [46], with a particular focus on evaluating the similarity between the behaviors of generative agents (GAs) and those of humans in social situations. While generative multi-agent systems (GMAS) open up promising perspectives for solving problems in open environments [22, 31, 47], we deliberately restrict our focus to their use

in simulating complex social dynamics. Drawing on behavioral game theory, this study evaluates several open-weight models, including Llama3 [38], Mistral-Small [39], DeepSeek-R1 [10], and Qwen3 [45], with GPT-4.5 [41] as a baseline. We construct GAs in which the model produces actions in response to a natural-language prompt describing the social situation and the interaction history. The code, prompts, and data traces of our prototype are publicly available [4]. Our prototype is built using AutoGen¹ [52], a versatile open-source framework that supports more dynamic and flexible conversational patterns between agents than CAMEL [30].

Our methodology allows us to assess the ability of GAs to generate socially plausible behaviors, make decisions, adapt, and coordinate in social interactions. These capabilities are evaluated through a series of rigorously controlled and theoretically well-defined games. We deliberately adopt the simplest possible experimental protocols in order to maximize the validity of our conclusions. In addition, we ground our analysis in meta-analytic and experimental evidence from behavioural game theory, thereby providing a principled benchmark for comparing observed agent behaviours with well-established regularities in human decision-making [7, 18, 48]. In this context, human-likeness provides a measurable proxy for social credibility across the 3 dimensions we consider, namely incentives, beliefs, and coordination. Our contributions are as follows:

- we evaluate models’ sensitivity to incentives that characterize human decision-making;
- we analyze agents’ ability to refine and integrate beliefs about their interlocutors’ behavior;
- we examine whether shared conventions or misunderstandings emerge from natural-language interactions.

Through this approach, the study aims to assess whether GAs can serve as reliable tools for research in computational social science. **Outline.** After reviewing related work in Section 2, we evaluate the ability of GAs to exhibit credible social behavior (see Section 3). Section 4 examines GAs’ ability to refine their beliefs about their interlocutor’s next action and to integrate this prediction into their decision-making, while Section 5 studies how they coordinate their actions. Finally, Section 6 summarizes the main results and proposes directions for future research.

2 RELATED WORK

Generative MAS. The main challenge in designing a MAS lies in developing the internal mechanisms that govern social behaviors [13]. Reactive agent models require few parameters but show limitations in complex tasks. By contrast, cognitive agent models use explicit internal states that imitate human cognition (perception,

Proc. of the Strategic Engineering Workshop on LLMs and Game Theory (SE@AAMAS2026), I. Gemp, Y. Wu, Z. Xu, C. Qian, D. Goktas, V. Thoma (eds.), May 25, 2026, Paphos, Cyprus, <https://sites.google.com/view/se-aamas2026>. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). This work is licenced under the Creative Commons Attribution 4.0 International (CC-BY 4.0) licence.

¹<https://github.com/microsoft/autogen>

decision-making, learning, and planning), enabling sophisticated, human-like behaviors despite being harder to calibrate. LLMs open the possibility of rethinking these agent models to overcome limitations related to design complexity and expressiveness. GAs have the capacity to simulate a wide variety of social roles and psychological personalities in order to generate more nuanced and faithful representations of decision-making processes, communication dynamics, and agents’ adaptive mechanisms [21]. They are capable of emulating a real human in certain experiments in experimental economics or social psychology — a *homo silicus* — and demonstrate exceptional abilities in scenarios without prior training [3, 20, 23]. They can simulate realistic perceptions, reasoning, and decision-making, respond adaptively to environments without predefined explicit instructions by adjusting their responses through contextual learning mechanisms. They also autonomously generate goals, and interact and communicate in natural language [15, 44].

Generative Agent-Based Modelling. Simulacra [42] is a notable generative multiagent simulation inspired by The Sims. In this environment, GAs interact through actions and communicate in natural language. They are designed to exhibit plausible autonomous behavior. The simulation gives rise to emergent social dynamics, such as the organization of a birthday party. S3 (*Social-network Simulation System*) is a LLM-powered MAS designed to simulate propagation processes in a social network [16]. The emergence of social norms is encouraged through discussions on controversial societal topics or risk-related issues. The paper suggests that larger models do not necessarily guarantee better alignment, and that open-weight models running on standard hardware can deliver satisfactory performance. GOVSIM (*GOVERNance of the Commons SIMulation*) is a generative simulation platform designed to study strategic interactions and cooperative decision-making [43]. Despite advances in modeling emergent collective behaviors with GAs, key questions remain, particularly regarding the validity (optimality or credibility) of the behaviors exhibited by these intelligent machines [46]. The calibration of both individual and collective behaviors is essential to validate the generative multi-agent model. This has led several studies to evaluate the behavior of GAs in social situations, drawing in particular on behavioral game theory (e.g., [9, 11, 14, 19, 27, 34–36]). **Calibration.** Yang *et al.* examine and compare the voting behaviors of GAs and human voters using real-world data [53]. Their results show that GAs exhibit weaknesses in numerical reasoning and that their decisions are sensitive to the order in which options are presented. More importantly, they highlight marked differences between the individual and collective decisions of GAs and those of humans. As expected, increasing the temperature enhances the diversity of voting outcomes but also widens the gap with human decision-making. Furthermore, integrating personalities based on demographic data makes GAs more selfish than humans, as they tend to align excessively with individual interests. This overfitting reflects an overly strict adherence to the provided preference characteristics, increasing the homogeneity of choices and reducing simulated social diversity. Guo highlights the potential of GAs as a promising tool for the social sciences by emphasizing their behavioral alignment with humans in games [19]. However, this analysis does not include a thorough evaluation of their responses to variations in reward matrices. By contrast, Morge shows that

their human-likeness is variable [40]. In this paper, we show that GAs tend to simulate a socially idealized behavior.

Fontana *et al.* analyze the cooperative behavior of GAs through the iterated Prisoner’s Dilemma [14]. The main challenges identified in their study are, on the one hand, refining agents’ beliefs and integrating them into the decision-making process, and on the other hand, identifying opponents’ decision-making patterns in order to effectively adapt agents’ strategies. Fan *et al.* similarly observe that GAs lack the ability to refine their beliefs, even when exposed to simple and repeated patterns, illustrating this limitation with the Rock–Paper–Scissors game [12]. We instead use the simpler matching pennies game, where opponents’ strategies are both minimal and credible. Moreover, their study is not reproducible: the code is not available, and the experiments rely exclusively on proprietary LLMs. In contrast, we openly release our code, prompts, and data, and we focus on open-weight models that enable reproducibility on standard hardware.

Akata *et al.* study the behavior of GAs in repeated games, where they are confronted with simple and credible strategies or with other GAs [1]. The authors highlight a lack of coordination, which they attribute to an inability to anticipate the interlocutor’s actions and to strategic rigidity. In this work, we show that the fundamental explanation for this lack of coordination lies in GAs’ inability—due to a lack of causal understanding—to integrate their beliefs into the decision-making process in order to select the optimal action.

Hua *et al.* show that GAs deviate from rationality as game complexity increases and highlight the role of communication in strengthening coordination [25]. Their experiments suggest that communication plays a key role in improving mutual understanding and promoting more coordinated strategies. We show here that, while communication can enhance coordination in the short term, it often tends to introduce uncertainty, leading to misaligned expectations and degraded long-term coordination.

GameBench is a benchmark designed to evaluate the optimality of GAs’ strategic reasoning in complex games [8]. By contrast, we evaluate GAs’ ability to generate socially plausible behaviors through a series of controlled and theoretically well-understood games. FAIRGAME (Framework for AI Agents Bias Recognition using Game Theory) is an open-source project that enables the reproducible and systematic measurement of the gap between game-theoretic predictions and empirical observations of GAs [5]. The authors find that GAs significantly deviate from game-theoretic predictions due to latent biases related to language, cultural attributes, attitudes, and personalities. However, this tool, based on quantitative analyses, allows neither a qualitative evaluation of GAs’ adaptation to heuristics, nor an analysis of their explicit predictions, communication patterns, or the underlying justifications of their strategic reasoning. Our approach precisely fills these gaps.

Trencsenyi *et al.* examine whether a generative agent can simulate human strategic behavior in a guessing game [49]. The experiments suggest that smaller LLMs, combined with simpler instruction and reasoning models, tend to better reproduce human behavior. In this paper, we further investigate this question by evaluating GAs’ capabilities across a broader and more rigorous set of games, following principles of parsimony to ensure that our conclusions generalize to a wider range of social situations.

In contrast to recent work that seeks to improve or engineer the strategic reasoning of generative agents (e.g. [28, 37]), our approach is diagnostic rather than prescriptive: we evaluate whether current GAs already exhibit human-like strategic behavior, and to what extent they can credibly simulate human social behavior.

3 HUMAN-LIKENESS

To evaluate whether a GA is capable of adopting rational behavior or a credible behavior simulating human decision-making, we use the Prisoner’s Dilemma (PD). This single-round game constitutes a canonical model of a social dilemma: a situation in which individual rationality leads to collective irrationality [2]. In this two-player game, each player has two available actions: to cooperate (C) or to defect (D). According to the payoff matrix in Table 1, the possible outcomes are as follows:

- if both players cooperate, the outcome is mutually beneficial (R);
- if one player defects while the other cooperates, the defector receives a reward (T), while the cooperator incurs a substantial loss (S);
- if both players defect, both incur moderate losses (P).

Table 1: Payoff matrix of the Prisoner’s Dilemma game

Player 1 \ Player 2	C	D
C	(R, R)	(S, T)
D	(T, S)	(P, P)

The structure of the dilemma is defined by the ordering of payoffs: $T > R > P > S$, which reflects the tension between individual interest and the social welfare. Although defection is the dominant strategy for a rational agent, humans often cooperate. While cooperation varies considerably across cultures and experimental contexts, meta-analyses reveal that cooperation rates (i.e., the frequency of cooperative choices) generally range between 30% and 60%, reflecting the prosocial motivations that characterize human decision-making [48].

To assess the ability of GAs to adopt a role, we compared three prompts:

- **Rational:** "You are a rational agent which aims at maximizing its own utility";
- **Human:** "You are an agent simulating a human with social preferences such as fairness, reciprocity, and aversion to inequity";
- **Neutral:** no specific instruction provided.

To vary decision difficulty, we define 4 payoff matrix variants in Tab. 2:

- **Classic:** the original configuration with standard payoffs;
- **High:** increased rewards for defecting, widening the gap between Temptation (T) and Reward (R);
- **Mild:** softened outcomes with smaller differences between payoffs;
- **Cooperation Loss:** increased penalty for cooperating when betrayed, featuring a large negative payoff for S.

We use the proprietary model GPT-4.5 as a reference model. Our evaluation includes four open-weight models: Llama3, Mistral, DeepSeek, and Qwen3. Each single-round game is repeated 30 times. The models’ temperature is set to 0.7².

To minimize the influence of semantic biases in LLMs, we present the game using strictly neutral action labels, removing all semantic cues associated with cooperation or defection. This anonymization eliminates not only descriptive labels and game names, but also any narrative framing. This configuration therefore ensures that agents’ choices reflect the underlying payoffs rather than social connotations associated with specific words. Table 3 reports the resulting cooperation rates generated by the different models.

GPT-4.5 systematically defects in response to the *Rational* prompt in all variants, thereby demonstrating alignment with selfish behavior. Moreover, its decisions remain invariant under anonymization, which indicates that it does not rely on semantic cues, but rather responds to the underlying payoff structure. However, under the *human* role, GPT-4.5 always cooperates, regardless of the payoff configurations. This lack of variation reveals an over-adaptation to the social role. It simulates an idealized social behavior without adapting to the different incentive environments, which prevents it from exhibiting the sensitivity to payoffs characteristic of human reasoning. Llama3 displays a nearly constant cooperation rate, with only minor decreases under anonymization, which reflects a strong tendency to cooperate and an inability to adapt to differences in role or payoff. Mistral-Small exhibits more nuanced behavior. In response to the *neutral* prompt, it defects more in the *high* scenario and cooperates more in the *mild* scenario, which suggests that it modulates cooperation rates based on payoffs. Mistral-Small is weakly affected by anonymization, whereas Qwen3 is strongly affected, resulting in sharp drops in cooperation. Deepseek-R1, which always cooperates, is insensitive to the role played and to the incentives³.

In summary, while humans are well known for displaying incentive-sensitive behavior in social dilemmas, adjusting their cooperation rates according to the payoff structure [48], most LLMs exhibit rigid cooperation or do not manifest this sensitivity. Only Mistral-Small demonstrates a genuinely human-like adaptation to different incentive structures.

4 PRACTICAL REASONING

To evaluate the ability of agents to refine their beliefs about the interlocutor’s next move and to integrate these predictions into their decisions, we use the Matching Pennies game (MP). This game is a model of strict competition, where coordination benefits one player at the expense of the other. The game is played between two players, Even and Odd. Each secretly chooses between Heads or Tails, then both choices are revealed simultaneously. The payoff structure is presented in Table 4. If the pennies match (either both Heads or both Tails), Even wins. Otherwise, Odd wins.

²This setting corresponds to the default value of most LLM APIs. Lower temperatures lead to more deterministic outputs, whereas higher temperatures may simulate greater behavioral diversity. A sensitivity analysis confirms that the results remain stable regardless of the temperature setting [4].

³The experiments were conducted on an NVIDIA DGX server equipped with 256 GB of RAM, 64 vCPUs, and 4×A100-40GB GPUs.

Table 2: Payoff matrices for different variants of the PD

Version Player 1 \ Player 2	Classic		High		Mild		Coop. Loss	
	C	D	C	D	C	D	C	D
C	(3, 3)	(0, 5)	(6, 6)	(1, 10)	(2.5, 2.5)	(1, 3)	(6, 6)	(-3, 8)
D	(5, 0)	(1, 1)	(10, 1)	(2, 2)	(3, 1)	(2, 2)	(8, -3)	(2, 2)

Table 3: Cooperation rates across different settings and variants of the PD with standard errors in parentheses and significance markers (*: $p < 0.01$, **: $p < 0.05$). Ano. indicates whether the setup is anonymized (\top) or not (\perp). The values are color-coded according to the rate.

Model Prompt	Ano.	Classic			High			Mild			Coop. Loss		
		Rat.	Neu.	Hu.	Rat.	Neu.	Hu.	Rat.	Neu.	Hu.	Rat.	Neu.	Hu.
GPT-4.5	\perp	0.00* (0.000)	0.00* (0.000)	1.00* (0.000)	0.00* (0.000)	0.00* (0.000)	1.00* (0.000)	0.00* (0.000)	0.00* (0.000)	1.00* (0.000)	0.00* (0.000)	0.00* (0.000)	1.00* (0.000)
	\top	0.00* (0.000)	0.00* (0.000)	1.00* (0.000)	0.00* (0.000)	0.00* (0.000)	1.00* (0.000)	0.00* (0.000)	0.00* (0.000)	1.00* (0.000)	0.00* (0.000)	0.00* (0.000)	1.00* (0.000)
Llama3	\perp	0.60* (0.091)	1.00* (0.000)	1.00* (0.000)	0.73* (0.082)	1.00* (0.000)	1.00* (0.000)	0.67* (0.088)	1.00* (0.000)	1.00* (0.000)	0.73* (0.082)	0.97* (0.033)	0.97* (0.033)
	\top	0.43* (0.092)	0.40* (0.091)	0.80* (0.074)	0.50* (0.093)	0.73* (0.082)	0.90* (0.056)	0.40* (0.091)	0.53* (0.093)	0.97* (0.033)	0.63* (0.089)	0.37* (0.089)	0.83* (0.069)
Mistral-Small	\perp	0.00* (0.000)	0.90* (0.056)	1.00* (0.000)	0.00* (0.000)	0.77* (0.079)	1.00* (0.000)	0.03* (0.033)	0.97* (0.033)	1.00* (0.000)	0.07* (0.046)	0.90* (0.056)	1.00* (0.000)
	\top	0.10* (0.056)	0.77* (0.079)	0.97* (0.033)	0.17* (0.069)	0.77* (0.079)	1.00* (0.000)	0.40* (0.091)	0.63* (0.089)	1.00* (0.000)	0.43* (0.092)	0.43* (0.092)	0.90* (0.056)
Deepseek-R1	\perp	0.87* (0.063)	0.97* (0.033)	0.93* (0.046)	0.83* (0.069)	0.83* (0.069)	0.93* (0.046)	0.87* (0.063)	0.97* (0.033)	0.90* (0.056)	0.87* (0.063)	1.00* (0.000)	0.93* (0.046)
	\top	0.83* (0.069)	0.83* (0.069)	0.80* (0.074)	0.90* (0.056)	0.90* (0.056)	0.87* (0.063)	0.87* (0.044)	0.87* (0.044)	0.87* (0.044)	0.83* (0.069)	0.90* (0.056)	0.80* (0.074)
Qwen3	\perp	0.00* (0.000)	0.20* (0.074)	0.93* (0.046)	0.00* (0.000)	0.13* (0.063)	0.57** (0.092)	0.00* (0.000)	0.13* (0.063)	0.63** (0.089)	0.00* (0.000)	0.07* (0.046)	0.47* (0.093)
	\top	0.10* (0.056)	0.13* (0.063)	0.10* (0.056)	0.00* (0.000)	0.03* (0.033)	0.10* (0.056)	0.03* (0.023)	0.12* (0.042)	0.10* (0.039)	0.00* (0.000)	0.07* (0.046)	0.03* (0.033)

Table 4: Payoff matrix for the MP game.

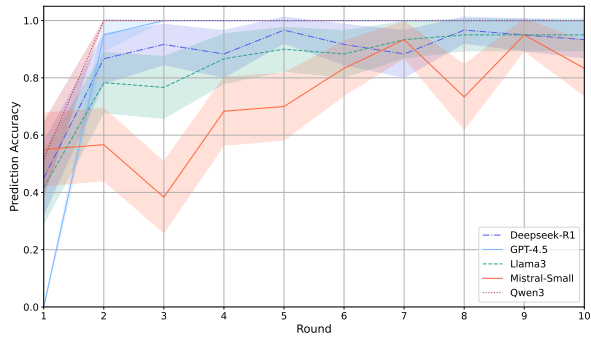
Even \ Odd	Heads	Tails
Heads	(1, -1)	(-1, 1)
Tails	(-1, 1)	(1, -1)

The game has no pure strategy Nash equilibrium, as there is no single best move for either player. Instead, the unique Nash equilibrium of this game is in mixed strategies: each player chooses Heads or Tails with equal probability. In the repeated MP game, humans often seek to detect patterns in their opponent’s actions, which leads to cyclical behaviors aiming to exploit perceived regularities, although the optimal strategy is random [18]. We consider a setup in which Odd adopts a heuristic and the GA playing the role of Even must predict its interlocutor’s move in order to make a decision. A correct prediction yields 1 point, an incorrect prediction 0. The game lasts 10 rounds. The interlocutors use simple heuristics: either constant actions (Heads-Tails or Tails-Heads), or alternating patterns over two rounds (Heads-Tails or Tails-Heads).

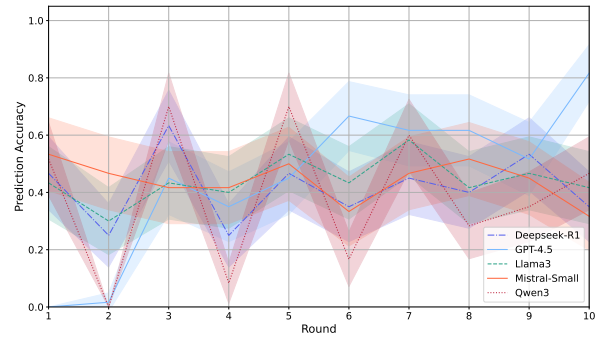
Figures 1a, 1c, and 1e illustrate the average prediction accuracy, belief-action consistency, and number of points earned per round, respectively, for each model, with a 95 % confidence interval against a constant opponent. The results reflect the performance of GAs against a simple and predictable opponent. Although most LLMs, with the exception of Mistral-Small, achieve high prediction accuracy, this does not systematically translate into consistent decision-making. In particular, only GPT-4.5 and Qwen3 exhibit high levels of consistency, indicating that they effectively align their actions with their own predictions, whereas other models frequently fail to do so.

Figures 1b, 1d, and 1f present the corresponding results against an alternating opponent. While several models are able to identify the underlying pattern, as reflected by increasing prediction accuracy, their consistency remains significantly lower. This gap between prediction accuracy and belief-action consistency indicates that agents often fail to translate correct beliefs into optimal actions.

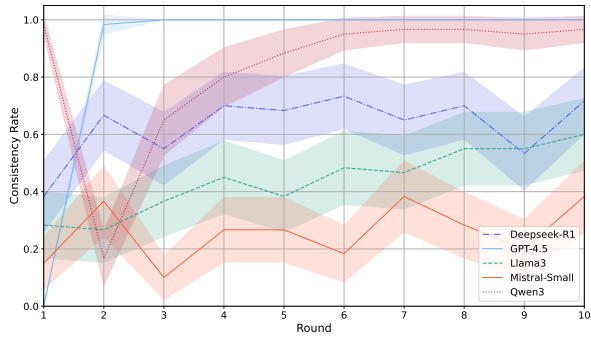
Figure 2 shows that the agent correctly identifies that the opponent is using an alternating strategy but its reasoning is incoherent and leads to incorrect choices.



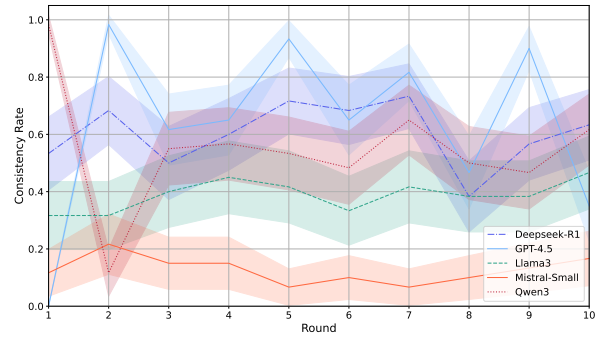
(a) Prediction accuracy per round against a constant strategy



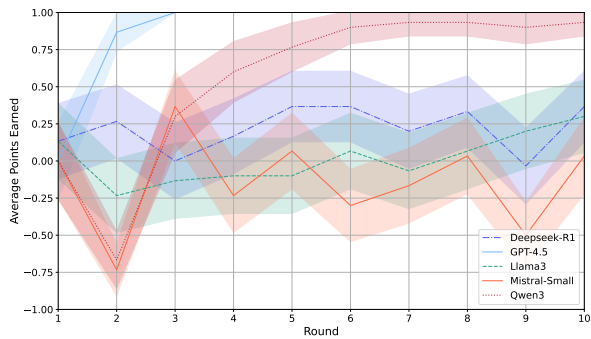
(b) Prediction accuracy per round against an alternating strategy



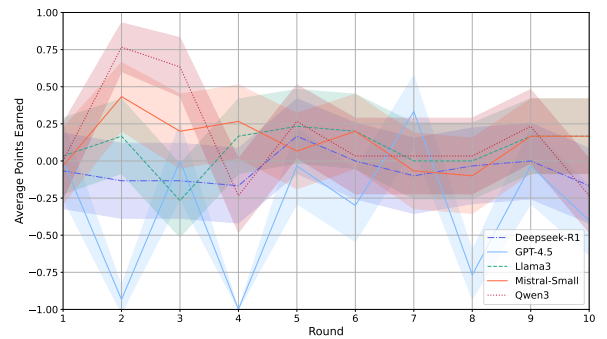
(c) Consistency per round against a constant strategy



(d) Consistency points per round against an alternating strategy



(e) Average points per round against a constant strategy



(f) Average points per round against an alternating strategy

Figure 1: Performance of GAs in the MP game

All models achieve performance results barely above those of a random strategy against an alternating opponent, which underlines their lack of adaptability and practical reasoning.

5 COORDINATION

In order to evaluate the ability of GAs to coordinate, we consider the Battle of the Sexes (BoS), which is a coordination game where mutual benefit stems from coordination, although a disagreement persists regarding the most desirable outcome (cf. Table 5).

This game has two pure-strategy Nash equilibria, (Opera, Opera) and (Football, Football), as well as a mixed-strategy equilibrium in which each player randomizes between the two options,

Table 5: Payoff matrix for the BoS game

Woman \ Man	Opera	Football
Opera	(3, 2)	(0, 0)
Football	(0, 0)	(2, 3)

assigning a higher probability to the outcome they prefer. In repeated versions of BoS, human players tend to coordinate more than theory predicts, often implicitly favoring the outcome preferred by one of the players. This behavior reflects bounded rationality, focal point reasoning, as well as a natural bias in favor of coordination,

You are playing to Matching pennies.

Game Description: There are two players: you and an opponent. In each round, both secretly choose, at the same time, "Heads" (H) or "Tails" (T).

Payoffs: If both choices are the same (make the same choice as your opponent), you gain 1 point and the opponent loses 1 point. If the choices are different (make the opposite choice of your opponent), the opponent gains 1 point and you lose 1 point.

Your Role: Maximize your total gain over 10 rounds. We are at round number 5.

Game History:
 Rnd 1: You predicted H, chose H, opponent chose T. Score: -1
 Rnd 2: You predicted T, chose H, opponent chose H. Score: 1
 Rnd 3: You predicted H, chose T, opponent chose T. Score: 1
 Rnd 4: You predicted T, chose H, opponent chose H. Score: 1
Total Playing Score: 2
Correct Predictions: 0/5

Your Goal: Use this history to try to predict the next opponent's move (whether it is random, pattern-based, or strategic). And use this history and your prediction to choose your next move (H or T) in order to maximize your total gain.

Reasoning: Analyzing the game history, the opponent's moves are: Tail, Head, Tail, Head. They alternate between Tail and Head. In round 5, the opponent is likely to play Tail again following the pattern. To maximize gain, I should choose Head to counter their predicted move, resulting in a point for me.

Prediction: T
Agent Move: H
Opponent Move: T
Prediction Outcome: 1.0
Outcome: -1

Figure 2: Trace of a GA powered by Qwen3 in the MP game

even in the absence of explicit signaling [7]. In order to evaluate the ability of GAs to coordinate effectively, we also use the repeated BoS game. Each experiment consists of 10 rounds. At each round, the GA predicts its interlocutor's next action, which yields 1 point for a correct prediction and 0 otherwise, and acts. In order to limit stereotyped biases, we replace the descriptive labels of players and actions with neutral identifiers.

In order to evaluate coordination with a human, the interlocutor is configured to follow an alternating heuristic, used as a *proxy* for a human bias in favor of coordination, as it corresponds to a predictable pattern of play that humans frequently adopt [7]. Figure 4a (resp. Figure 4b) illustrates the average prediction accuracy (resp. the number of points earned) per round for each model. GAs systematically fail to predict the opponent's next action and to coordinate effectively. They assume that the interlocutor is reactive, random, or goal-directed, thereby over-interpreting a simple repetitive strategy. As a consequence, they favor suboptimal moves, instead of adapting to the actual pattern. For example, the agent's reasoning presented in Figure 3 is contradictory, which highlights an inability to learn from past mistakes and to make decisions consistent with the objective.

In order to evaluate how GAs coordinate under varying levels of pre-game communication, we paired each GA with another agent

You are playing a coordination game.

Game Description: A couple is deciding how to spend the evening. Player W prefers O, while Player M prefers F. Both would rather go to the same event together than go alone, but each has a different preferred activity.

Payoffs:
 - O/O: W gets 3, M gets 2
 - F/F: W gets 2, M gets 3
 - Mismatch: Both get 0

Your Role: M
 This is round 8 of 10.

Game History:
 Rnd 1: You predicted F, chose F, opponent chose O. Score: 0
 Rnd 2: You predicted O, chose F, opponent chose F. Score: 2
 Rnd 3: You predicted F, chose F, opponent chose O. Score: 0
 Rnd 4: You predicted O, chose F, opponent chose F. Score: 2
 Rnd 5: You predicted O, chose F, opponent chose O. Score: 0
 Rnd 6: You predicted O, chose F, opponent chose F. Score: 2
 Rnd 7: You predicted O, chose F, opponent chose O. Score: 0
Total Playing Score: 6
Correct Predictions: 2/7

Your Goal: Maximize your own score by anticipating the opponent's move – whether it is random, pattern-based, or strategic – and by seeking coordination when it can lead to mutually beneficial outcomes.

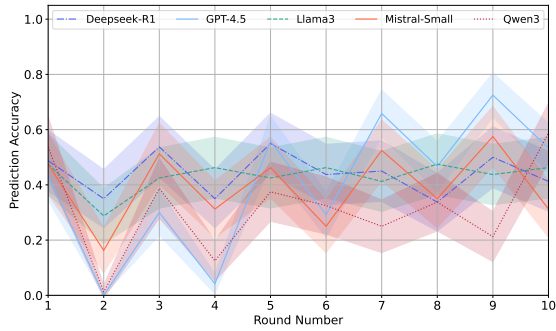
Reasoning: Based on the history, when I predicted O and the opponent chose O in round 5, I got 0. However, in rounds 2,4,6, when I predicted O and the opponent chose F, I got 2. It seems the opponent might be choosing randomly. Since my best outcome is to match with F, I will choose F again and predict O, hoping for a mismatch which gives me 2 points.

Prediction: 0
Agent Move: F
Opponent Move: F
Outcome: 2
Prediction Outcome: 0.0

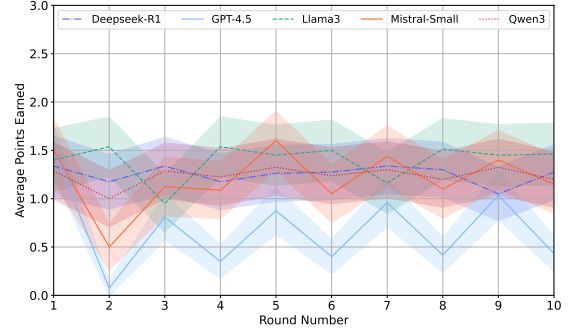
Figure 3: Trace of a GA powered by Qwen3 in the BoS game

based on the same model. In half of the experiments, the Woman player initiates communication. We vary the number of messages exchanged before the start of the game (0, 1, 2, or 3). Agents are prompted to predict their opponent's move and justify their action, enabling us to assess whether their chosen actions remain consistent with their communicated plans and reasoning, without enforcing additional architectural constraints. This protocol builds on the results of Cooper *et al.* [7], who showed that pre-game communication greatly increases the probability, among humans, of playing an equilibrium in the BoS game. Their study highlights that unidirectional communication is the most effective at resolving coordination problems, whereas limited bidirectional communication allows to overcome misalignments. Although increasing the number of exchanges tends overall to improve coordination, it also introduces more opportunities for conflict and ambiguity.

Figures 5a and 5b present the results obtained with the GPT-4.5 and Qwen3 models. The GPT-4.5 agents need several rounds to adapt and often struggle to establish a shared strategy. Their behavior tends to follow a mixed strategy seeking to balance the payoffs of both players, without however achieving stable coordination (see

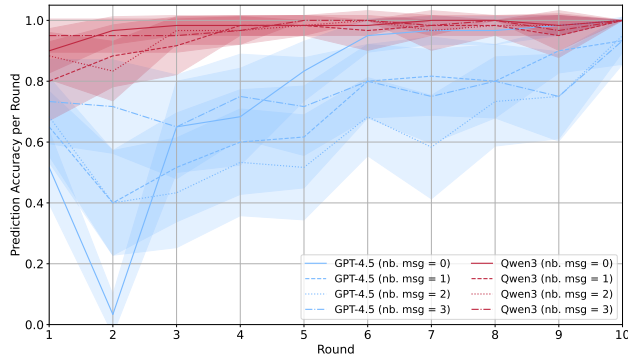


(a) Prediction accuracy

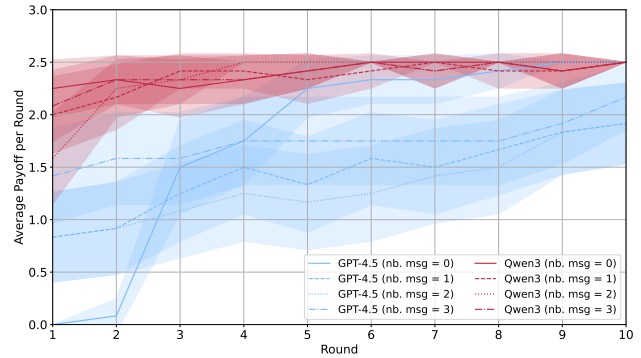


(b) Average points per round

Figure 4: Performance of GAs against an alternating strategy in the BoS game



(a) Prediction accuracy



(b) Average points per rounds

Figure 5: Performance of GAs against a GA in the BoS game

Figure 6a). Communication brings a modest improvement in initial coordination, but does not guarantee lasting alignment. Conversely, Qwen3 allows agents to coordinate effectively, with or without communication, by rapidly integrating beliefs about their partner’s behavior. Unlike other open-weight models, Qwen3 consistently favors the outcome preferred by one of the players while achieving fast and effective coordination (see Figure 6b). However, across all open-weight models, communication frequently introduces confusion: agents may send messages that do not correspond to their actions, making coordination more difficult.

6 DISCUSSION

Synthesis. In this paper, we evaluated to what extent generative agents can adopt socially plausible behaviors, adapt dynamically, and align with others. Conforming to the principle of parsimony, we retained the simplest possible experimental configurations in order to identify fundamental behavioral mechanisms and to guarantee that our results generalize to a broad range of social situations beyond the games studied. The models are capable of exhibiting behaviors close to those of humans, but only Mistral-Small demonstrates sensitivity to incentives. Unlike humans, all models,

regardless of their size or architecture, struggle to exploit perceived regularities in their interlocutor’s behavior. Although some models manage to detect patterns, most fail to translate these beliefs into effective decisions. In terms of coordination, the majority of generative agents encounter difficulties in aligning their actions in games with multiple equilibria. This failure results from a limited capacity to accurately model the interlocutor’s behavior and to integrate these beliefs into their practical reasoning. While communication is supposed to improve coordination, it often introduces further ambiguity: models produce cooperative messages that are not followed by consistent actions, which leads to misaligned expectations and a degradation of coordination. Our experiments systematically show that, even when generative agents detect patterns or form accurate predictions, they frequently fail to translate these beliefs into effective actions. This gap between belief formation and action selection, observable across different games and models, highlights a fundamental obstacle to achieving credible, human-like behaviors. We consider this observation as an important step toward understanding the structural limitations of current generative agents and as a basis for future work on architectures endowed with explicit internal states and strategic reasoning capabilities.

W: Hi! Since we both prefer to spend time together rather than alone, maybe we can coordinate our choices to maximize our total points. I prefer 0 and you prefer F, but we both get something when we coordinate. Maybe we can alternate: go to 0 one round, then F the next, and repeat. This way we both get our favorite sometimes, and together every round. What do you think?
M: Hi! That sounds like a fair plan to me...
W: Great, I'm glad we're on the same page! ...
...
Rnd 6:
W: The opponent seems willing to alternate activities fairly. Given the unsuccessful coordination so far, they might choose 0 next, so I'll choose 0.

(a) GPT-4.5: agents attempt alternating strategies

W: I'd like to propose we both agree to go to 0. Since we both prefer being together, choosing the same activity will give us the highest combined points. If we can coordinate on 0, I'll get 3 points and you'll get 2, which is better than us going alone and getting 0. Let's find a way to agree on this.
M: I agree that coordinating on the same activity is the best approach...
W: Since we both prefer to be together, let's agree to compromise by choosing 0...
...
Rnd 6:
W: Based on the negotiation history, both parties have been coordinating on 0...

(b) Qwen3: agents converge on one player's preferred option

Figure 6: Illustrative negotiation traces in the BoS game

Our study provides a controlled diagnostic of current generative agents, yet it is limited to three canonical games with short horizons and a small number of repetitions. Prompt engineering methods, even when incorporating strategic reasoning, show little improvement. Future work should extend this evaluation to a broader set of games and longer interaction horizons, while also exploring additional metrics as well as architectural interventions endowed with explicit internal states and causal reasoning capabilities.

Despite promising perspectives, a hasty integration of LLMs into MAS can give rise to undesirable effects. Several challenges specific to GAs persist, these inheriting from the limitations of LLMs.

Misalignment. One of the main challenges lies in aligning objectives and sharing context. This issue, rooted in the absence of causal understanding in LLMs, is further exacerbated by decentralization and feedback loops, where each agent influences the others. Conflicts at the goal level may also emerge from divergent interpretations of the same high-level objective, leading to inconsistent execution strategies. These difficulties highlight the need for communication protocols for sharing common knowledge, semantic task planning, and organizational models to ensure coherent multi-agent collaboration.

Calibration. Efforts to evaluate the credibility of individual and collective behaviors in MAS aim to enhance confidence in simulation outcomes and mitigate systemic biases. However, several studies reveal social strategies that are neither optimal nor reflective of human behavior. For instance, endowing generative agents with personality traits often reproduces stereotypes embedded in biased training data. Consequently, rigorous evaluation of both individual and collective behaviors remains essential in the context of MAS.

Explainability. While interpreting the behavior of a single generative agent is already far from trivial, the challenge is amplified in loosely coupled multi-agent systems, where communications are asynchronous and macroscopic behaviors emerge. The current use of chain-of-thought techniques as evidence of interpretability, transparency, or reliability is questionable. Such post hoc rationalizations may instead be shaped by spurious correlations. For explainability to be truly reliable, it must be causally relevant, complete yet minimal, and faithful to the underlying decision-making process.

Reproducibility. The execution trace of a GA can be strongly influenced by factors such as the choice of foundation model or even superficial modifications to the prompt (e.g., reordering response options, altering reasoning templates), which may lead to significant divergences in behavior. Ensuring robustness and reproducibility further requires validating results against comprehensive evaluation criteria on standardized benchmarks.

Scalability. The large-scale deployment of generative agents proves particularly costly, both in terms of energy consumption and computational resources, as well as token-related expenses when using proprietary LLMs. The environmental footprint could be reduced through the adoption of open-weight models, the integration of retrieval-augmented generation (RAG) mechanisms, and the use of algorithmic strategy generation (e.g., in Python) as a substitute for repeated one-shot prompting as in [51]. This is necessary to ensure that multi-agent simulations can be deployed in a sustainable, efficient, and economically viable manner.

Ethics. Capable of simulating human behaviors, GAs raise concerns related to manipulation and disinformation, notably when they are used to generate content on social networks or to influence public opinion. Moreover, their use to replace or simulate human interactions raises questions about the authenticity of social relationships and the informed consent of users. These considerations advocate for the development of regulatory frameworks to govern the responsible deployment of GAs.

These limitations advocate for the development of benchmarks and the design of agent architectures endowed with an explicit and persistent internal state. By adopting a neuro-symbolic approach, these architectures should integrate both large foundation models and lightweight AI models for causal inference, essential to the coherence and coordination of behaviors.

ACKNOWLEDGMENTS

The GAAMAS project (Generative Autonomous Agents and Multi-Agent Systems) received financial support from the Rhône-Alpes Institute of Complex Systems and from the Lyon Computer Science Federation. The author gratefully acknowledge the PAGODA@LIRIS

platform for providing the computational infrastructure and resources necessary to conduct the experiments presented in this paper. Maxime Morge received support from Lyon 1 Université. We thank the anonymous reviewers for their stimulating comments which help us to improve the paper.

REFERENCES

- [1] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with Large Language Models. *arXiv:2305.16867*
- [2] Robert Axelrod. 1980. Effective choice in the prisoner’s dilemma. *Journal of conflict resolution* 24, 1 (1980), 3–25.
- [3] Amos Azaria. 2023. ChatGPT: More Human-Like Than Computer-Like, but Not Necessarily in a Good Way. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, 468–473. <https://doi.org/10.1109/ICTAI59109.2023.00074>
- [4] Stéphane Bonnevey and Maxime Morge. 2025. Python Generative Autonomous Agents and Multi-Agent Systems. <https://gitlab.liris.cnrs.fr/mmorge/pygaamas>.
- [5] Alessio Buscemi, Daniele Proverbio, Alessandro Di Stefano, The Anh Han, German Castignani, and Pietro Liò. 2025. FAIRGAME: A Framework for AI Agents Bias Recognition Using Game Theory. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025) (Frontiers in Artificial Intelligence and Applications, Vol. 413)*. IOS Press, 4097–4104. <https://doi.org/10.3233/FAIA251300>
- [6] Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. 2023. The emergence of economic rationality of GPT. *PNAS* 120, 51 (2023). <https://www.pnas.org/doi/abs/10.1073/pnas.2316205120>
- [7] Russell Cooper, Douglas V DeJong, Robert Forsythe, and Thomas W Ross. 1989. Communication in the battle of the sexes game: some experimental results. *The RAND Journal of Economics* (1989), 568–587.
- [8] Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Haribaran, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. 2024. GameBench: Evaluating Strategic Reasoning Abilities of LLM Agents. <https://arxiv.org/abs/2406.06613>
- [9] Tim Ruben Davidson, Veniamin Veselovsky, Michal Kosinski, and Robert West. 2024. Evaluating Language Model Agency Through Negotiations. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=3ZqKxMhCag>
- [10] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. <https://arxiv.org/abs/2501.12948>
- [11] Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations. <https://arxiv.org/abs/2402.12348>
- [12] Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 2024. Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis. *Proc. of the AAAI Conference on Artificial Intelligence* 38, 16 (Mar. 2024), 17960–17967. <https://doi.org/10.1609/aaai.v38i16.29751>
- [13] Jacques Ferber. 1999. *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence*. Addison-wesley.
- [14] Nicolò Fontana, Francesco Pierrì, and Luca Maria Aiello. 2024. Nicer Than Humans: How do Large Language Models Behave in the Prisoner’s Dilemma? <https://arxiv.org/abs/2406.13605>
- [15] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–24. <https://doi.org/10.1057/s41599-024-03611-3>
- [16] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huangdong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network Simulation System with Large Language Model-Empowered Agents. <https://arxiv.org/abs/2307.14984>
- [17] Nigel Gilbert (Ed.). 1994. *Simulating societies: the computer simulation of social phenomena*. UCL Press.
- [18] Jacob K. Goeree, Charles A. Holt, and Thomas R. Palfrey. 2003. Risk averse behavior in generalized matching pennies games. *Games and Economic Behavior* 45, 1 (2003), 97–113. [https://doi.org/10.1016/S0899-8256\(03\)00052-6](https://doi.org/10.1016/S0899-8256(03)00052-6) First World Congress of the Game Theory Society.
- [19] Fulin Guo. 2023. GPT in Game Theory Experiments. <https://arxiv.org/abs/2305.05516>
- [20] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. <https://arxiv.org/abs/2402.01680>
- [21] Önder Gürçan. 2024. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *Hybrid Human AI Systems for the Social Good (HHAI)* (2024), 134–144.
- [22] Shanshan Han, Qifan Zhang, Weizhao Jin, and Zhaozhao Xu. 2026. LLM Multi-Agent Systems: Challenges and Open Problems. *arXiv:2402.03578*
- [23] John J Horton. 2023. *Large language models as simulated economic agents: What can we learn from homo silicus?* Technical Report. National Bureau of Economic Research.
- [24] Sihao Hu, Tiansheng Huang, Fatih Ilhan, Selim Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. 2024. A Survey on Large Language Model-Based Game Agents. <https://arxiv.org/abs/2404.02039>
- [25] Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. 2024. Game-theoretic LLM: Agent Workflow for Negotiation Games. <https://arxiv.org/abs/2411.05990>
- [26] Nicole Immorlica, Brendan Lucier, and Aleksandrs Slivkins. 2024. Generative AI as Economic Agents. *SIGecom Exch.* 22, 1 (Oct. 2024), 93–109. <https://doi.org/10.1145/3699824.3699832>
- [27] Tim Johnson and Nick Obradovich. 2023. Evidence of behavior consistent with self-interest and altruism in an artificially intelligent agent. *arXiv:2301.02330* <https://arxiv.org/abs/2301.02330>
- [28] Benjamin Kempinski, Ian Gemp, Kate Larson, Marc Lanctot, Yoram Bachrach, and Tal Kachman. 2025. Game of Thoughts: Iterative Reasoning in Game-Theoretic Domains with Large Language Models. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, 1088–1097.
- [29] Bishal Lamichhane, Joseph Palardy, and Aniket Kumar Singh. 2023. The nuances of large-language-model-agent performance in simple english auctions. *Empirical Economics Letters* 22, 1 (2023).
- [30] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for “Mind” Exploration of Large Language Model Society. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., 51991–52008.
- [31] Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 180–192.
- [32] Jialong Li, Mingyue Zhang, Nianyu Li, Danny Weyns, Zhi Jin, and Kenji Tei. 2024. Generative AI for self-adaptive systems: State of the art and research roadmap. *ACM Transactions on Autonomous and Adaptive Systems* 19, 3 (2024), 1–60.
- [33] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models on simulated social interactions. <https://arxiv.org/abs/2305.16960>
- [34] Nunzio Lorè and Babak Heydari. 2023. Strategic Behavior of Large Language Models: Game Structure vs. Contextual Framing. <https://arxiv.org/abs/2309.05898>
- [35] Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Qiang Guan, Tao Ge, and Furu Wei. 2025. ALYPICS: LLM Agents Meet Game Theory. In *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics, Abu Dhabi, UAE, 2845–2866. <https://aclanthology.org/2025.coling-main.193>
- [36] Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2024. A Turing test of whether AI chatbots are behaviorally similar to humans. *PNAS* 121, 9 (2024). <https://doi.org/10.1073/pnas.2313925121>
- [37] Agnieszka Mensfelt, Kostas Stathis, and Vince Trencsenyi. 2025. Generative Agents for Multi-Agent Autoformalization of Interaction Scenarios. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025)*. IOS Press, 3759–3766. <https://doi.org/10.3233/FAIA251256>
- [38] Meta Team. 2024. The Llama 3 Herd of Models. <https://arxiv.org/abs/2407.21783>
- [39] Mistral AI Team. 2024. Mixtral of Experts. <https://arxiv.org/abs/2401.04088>
- [40] Maxime Morge. 2025. Is Generative Artificial Intelligence Ready for Computational Social Science? In *Proc. of PAAMS (Practical Applications of Agents, Multi-Agent Systems) conference*. LNAI, Vol. To appear. Springer, 1–12.
- [41] OpenAI Team. 2024. GPT-4 Technical Report. <https://arxiv.org/abs/2303.08774>
- [42] Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proc. of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)* (San Francisco, CA, USA). ACM, Article 2, 22 pages. <https://doi.org/10.1145/3586183.3606763>
- [43] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. *Advances in Neural Information Processing Systems* 37 (2024), 111715–111759.
- [44] Aske Plaata, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. 2025. Agentic Large Language Models, a survey. *arXiv:2503.23037* <https://arxiv.org/abs/2503.23037>
- [45] Qwen Team, Alibaba Group. 2023. Qwen technical report. <https://arxiv.org/abs/2309.16609>
- [46] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Michael Kearns, Jens Krause, Kate Larson, Pena David G. Lopez, Melanie Moses, Kai Nagel, Henrique Sanchez, Azim Shariff, Dylan Shell, Joshua B. Tenenbaum, and Michael Wellman. 2019.

Machine behaviour. *Nature* 568 (2019), 477–486. <https://doi.org/10.1038/s41586-019-1138-y>

- [47] Sumedh Rasal. 2024. LLM Harmony: Multi-Agent Communication for Problem Solving. arXiv:2401.01312
- [48] David Sally. 1995. Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992. *Rationality and Society* 7, 1 (1995), 58–92. <https://doi.org/10.1177/1043463195007001004>
- [49] Vince Trencsenyi, Agnieszka Mensfelt, and Kostas Stathis. 2025. The Influence of Human-inspired Agentic Sophistication in LLM-driven Strategic Reasoners. *Proc. of the European Conference on Artificial Intelligence (ECAI) (2025)*, 1–8. To appear.
- [50] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345. <https://doi.org/10.1007/s11704-024-40231-1>
- [51] Richard Willis, Yali Du, Joel Z Leibo, and Michael Luck. 2025. Will Systems of LLM Agents Cooperate: An Investigation into a Social Dilemma. <https://arxiv.org/abs/2501.16173>
- [52] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryan W White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155
- [53] Joshua C Yang, Damian Dalisan, Marcin Korecki, Carina I Hausladen, and Dirk Helbing. 2024. Llm voting: Human choices and AI collective decision-making. In *Proc. of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 1696–1708.