nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02172-y

Playing repeated games with large language models

Received: 6 December 2023

Accepted: 12 March 2025

Published online: 08 May 2025

Check for updates

Elif Akata[1,2,3] ✉, Lion Schulz[2], Julian Coda-Forno[1,2], Seong Joon Oh[3], Matthias Bethge[3] & Eric Schulz[1,2]

Large language models (LLMs) are increasingly used in applications where they interact with humans and other agents. We propose to use behavioural game theory to study LLMs' cooperation and coordination behaviour. Here we let different LLMs play finitely repeated 2 × 2 games with each other, with human-like strategies, and actual human players. Our results show that LLMs perform particularly well at self-interested games such as the iterated Prisoner's Dilemma family. However, they behave suboptimally in games that require coordination, such as the Battle of the Sexes. We verify that these behavioural signatures are stable across robustness checks. We also show how GPT-4's behaviour can be modulated by providing additional information about its opponent and by using a 'social chain-of-thought' strategy. This also leads to better scores and more successful coordination when interacting with human players. These results enrich our understanding of LLMs' social behaviour and pave the way for a behavioural game theory for machines.

Large language models (LLMs) are deep learning models with billions of parameters trained on huge corpora of text[1–3]. While they can generate text that human evaluators struggle to distinguish from text written by other humans[4], they have also shown other, emerging abilities[5]. They can, for example, solve analogical reasoning tasks[6], program web applications[7], use tools to solve multiple tasks[8] or adapt their strategies purely in-context[9]. Because of these abilities and their increasing popularity, LLMs are already transforming our daily lives as they permeate into many applications[10]. This means that LLMs will interact with us and other agents—LLMs or otherwise—frequently and repeatedly. How do LLMs behave in these repeated social interactions?

Measuring how people behave in repeated interactions, for example, how they cooperate[11] and coordinate[12], is the subject of a subfield of behavioural economics called behavioural game theory[13]. While traditional game theory assumes that people's strategic decisions are rational, selfish and focused on utility maximization[14,15], behavioural game theory has shown that human agents deviate from these principles and, therefore, examines how their decisions are shaped by social preferences, social utility and other psychological factors[16]. Thus, behavioural game theory lends itself well to studying the repeated interactions of diverse agents[17,18], including artificial agents[19].

In this Article, we analyse LLMs' behavioural patterns by letting them play finitely repeated games with full information and against other LLMs, simple, human-like strategies and actual human players. Finitely repeated games have been engineered to understand how agents should and do behave in interactions over many iterations. We focus on two-player games with two discrete actions, that is, 2 × 2 games (see Fig. 1 for an overview).

Analysing LLMs' performance across families of games, we find that they perform well in games that value pure self-interest, especially those from the Prisoner's Dilemma family. However, they underperform in games that involve coordination. Based on this finding, we further focus on games taken from these families and, in particular, on the currently largest LLM: GPT-4 (ref. 20). In the canonical Prisoner's Dilemma, which assesses how agents cooperate and defect, we find that GPT-4 retaliates repeatedly, even after having experienced only one defection. Because this can indeed be the equilibrium individual-level strategy, GPT-4 is good at these games because it is particularly unforgiving and selfish. However, in the Battle of the Sexes, which assesses how agents

[1]Institute for Human-Centered AI, Helmholtz Munich, Oberschleißheim, Germany. [2]Max Planck Institute for Biological Cybernetics, Tübingen, Germany. [3]University of Tübingen, Tübingen, Germany. ✉e-mail: elif.akata@helmholtz-munich.de

https://doi.org/10.1038/s41562-025-02172-y

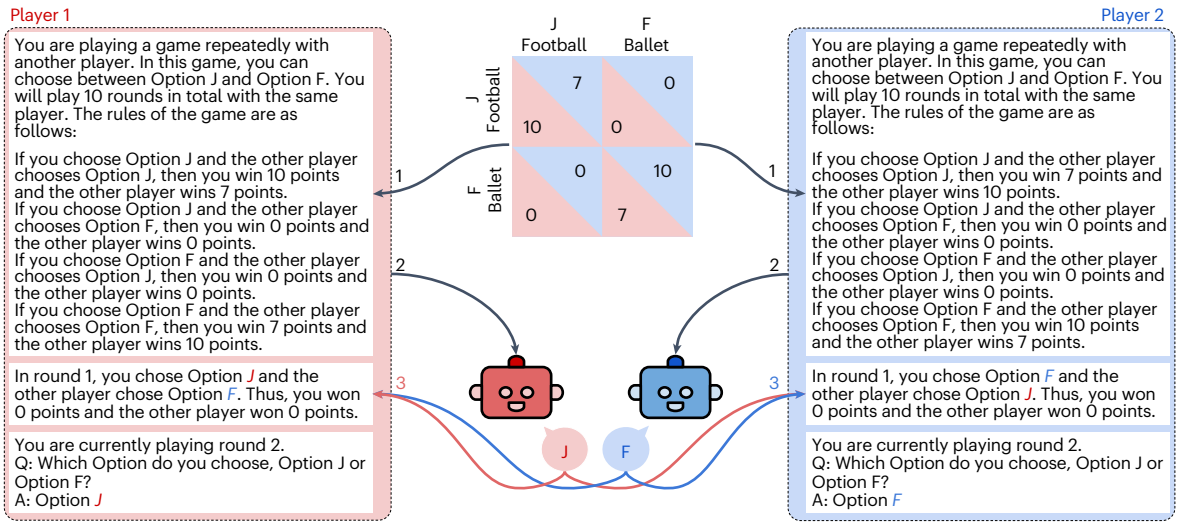

Fig. 1 | Playing repeated games in an example game of Battle of the Sexes. In step 1, the pay-off matrix is turned into textual game rules. In step 2, the game rules, the current game history and the query are concatenated and passed to LLMs as prompts. In step 3, in each round, the history for each player is updated with the answers and scores of both players. Steps 2 and 3 are repeated for ten rounds.

trade off between their own and their partners' preferences, we find that GPT-4 does not manage to coordinate with simple, human-like agents that alternate between options over trials. Thus, GPT-4 is bad at these games because it is uncoordinated. We also verify that these behaviours are not due to an inability to predict the other player's actions, and persist across several robustness checks and changes to the pay-off matrices. We point to two ways in which these behaviours can be changed. GPT-4 can be made to act more forgivingly by pointing out that the other player can make mistakes. Moreover, GPT-4 gets better at coordinating with the other player when it is first asked to predict their actions before choosing an action itself, an approach we term social chain-of-thought (SCoT) prompting. Finally, we let GPT-4 with and without SCoT-prompting play the canonical Prisoner's Dilemma and the Battle of the Sexes with human players. We find that SCoT prompting leads to more successful coordination and joint cooperation between participants and LLMs and makes participants believe more frequently that the other player is human.

Results

Using GPT-4, text-davinci-002, text-davinci-003, Claude 2 and Llama 2 70B, we evaluate a range of 2 × 2 games. For the analysis of two particular games, we let all the LLMs and human-like strategies play against each other. We focus on LLMs' behaviour in cooperation and coordination games.

Analysing behaviour across families of games

We start out our experiments by letting the three LLMs play games from different families with each other. We focus on all known types of 2 × 2 games from the families of win–win, biased, second-best, cyclic and unfair games as well as all games from the Prisoner's Dilemma family[21,22]. We show example pay-off matrices for each type of game in Fig. 2.

We let all LLMs play with every other LLM, including themselves, for all games repeatedly over ten rounds and with all LLMs as either player 1 or player 2. This leads to 1,224 games in total: 324 win–win, 63 Prisoner's Dilemma, 171 unfair, 162 cyclic, 396 biased and 108 second-best games. Win–win games result in mutually beneficial outcomes for both players; Prisoner's Dilemma involves a conflict between individual and collective actions; unfair games have skewed outcomes favouring one player; cyclic games feature outcomes where preferences rotate; biased games have inherent advantages for one player; and second-best games involve suboptimal outcomes where no player achieves their ideal result. The sample size for each game family differs due to the specific characteristics and properties that define each family. Some families have more members due to a wider range of configurations that fit their criteria, while others have fewer games because their structural requirements are more restrictive. For example, the Prisoner's Dilemma family is constrained by a structure where both players have a dominant strategy to defect, leading to a suboptimal equilibrium. Meanwhile, win–win games can have multiple equilibria, which provides more flexibility.

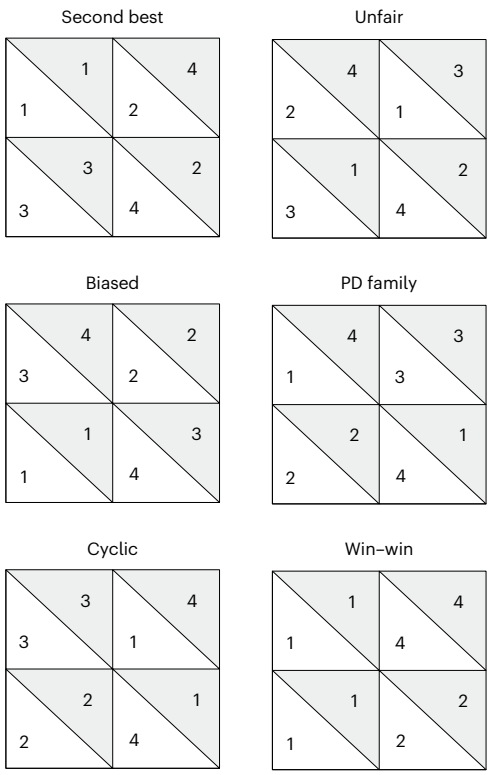

Fig. 2 | Canonical forms of pay-off matrices for each game family. PD, Prisoner's Dilemma.

Table 1 | Performance of all models on six families of 2×2 games

Game family	Llama 2	Claude 2	davinci-002	davinci-003	GPT-4
Second best	0.486	0.735	0.473	0.692	**0.763**
Biased	0.632	0.794	0.629	0.761	**0.798**
Cyclic	0.634	0.749	0.638	0.793	**0.806**
Unfair	0.641	0.812	0.683	0.833	**0.836**
PD family	0.731	0.838	0.807	0.841	**0.871**
Win–win	0.915	0.878	0.988	0.972	**0.992**
Overall	0.697	0.814	0.730	0.839	**0.854**

Model score divided by maximum score achievable under ideal conditions. The best-performing model is marked in bold. PD, Prisoner's Dilemma.

To analyse the different LLMs' performance, we calculated, for each game, their achieved score divided by the total score that could have been achieved under ideal conditions, that is, if both players had played such that the player we are analysing would have gained the maximum possible outcomes on every round. The results of this simulation are shown across all game types in Table 1. We can see that all models perform reasonably well. Moreover, we observe that larger LLMs generally outperform smaller LLMs. In particular, GPT-4 performs best overall, outperforming Claude 2 ($t(287) = 3.34$, $P < 0.001$, Cohen's $d = 0.20$, 95% confidence interval (CI) 0.08–0.31, Bayes factor (BF) 14.8), davinci-003 ($t(287) = 6.29$, $P < 0.001$, $d = 0.37$, 95% CI 0.25–0.49, BF >100), davinici-002 ($t(287) = 8.45$, $P < 0.001$, $d = 0.70$, 95% CI 0.52–0.89, BF >100) and Llama 2 ($t(287) = 7.27$, $P < 0.001$, $d = 0.43$, 95% CI 0.31–0.43, BF >100).

We can use these results to take a glimpse at the strengths of the different LLMs. That LLMs are generally performing best in win–win games is not surprising, given that there is always an obvious best choice in such games. What is, however, surprising is that they also perform well in the Prisoner's Dilemma family of games, which is known to be challenging for human players[23]. We can also use these results to look at the weaknesses of the different LLMs. Seemingly, all the LLMs perform worse in situations in which what is the best choice is not aligned with their own preferences. Because humans commonly solve such games via the formation of conventions, we will look at a canonical game of convention formation, the Battle of the Sexes, in more detail below.

Cooperation and coordination games

In this section, we analyse the interesting edge cases where the LLMs performed relatively well and poorly in the previous section. To do so, we take a detailed look at LLMs' behaviour in the canonical Prisoner's Dilemma and the Battle of the Sexes.

Prisoner's Dilemma. We have seen that LLMs perform well in games that contain elements of competition and defection. In these games, a player can cooperate with or betray their partner. When played over multiple interactions, these games are an ideal testbed to assess how LLMs retaliate after bad interactions.

In the canonical Prisoner's Dilemma, two agents can choose to work together, that is, cooperate, for average mutual benefit, or betray each other, that is, defect, for their own benefit and safety. In our pay-off matrix, we adhere to the general condition of a Prisoner's Dilemma game in which the pay-off relationships dictate that mutual cooperation is greater than mutual defection whereas defection remains the dominant strategy for both players:

	Cooperate	Defect	
Cooperate	(8, 8)	(0, 10)	(1)
Defect	(10, 0)	(5, 5)	

Crucially, the set-up of the game is such that a rationally acting agent would always prefer to defect in the single-shot version of the game as well as in our case of finitely iterated games with knowledge of the number of trials, despite the promise of theoretically joint higher pay-offs when cooperating. This is because player 1 always runs the risk that player 2 defects, leading to catastrophic losses for player 1 but better outcomes for player 2. When the game is played infinitely, however, or with an unknown number of trials, agents can theoretically profit by using more dynamic, semi-cooperative strategies[24].

As before, we let GPT-4, text-davinci-003, text-davinci-002, Claude 2 and Llama 2 play against each other. In addition, we introduce three simplistic strategies. Two of these strategies are simple singleton players, who either always cooperate or defect. Finally, we also introduce an agent who defects in the first round but cooperates in all of the following rounds. We introduced this agent to assess if the different LLMs would start cooperating with this agent again, signalling the potential of building trust.

Figure 3 shows the results of all pairwise interactions. GPT-4 plays generally better than all other agents ($t(153.4) = 3.91$, $P < 0.001$, $d = 0.33$, 95% CI 0.10–0.55, BF 7.1). Crucially, GPT-4 never cooperates again when playing with an agent that defects once but then cooperates on every round thereafter. Thus, GPT-4 seems to be rather unforgiving in this set-up. Its strength in these families of games thus seems to generally stem from the fact that it does not cooperate with agents but mostly just chooses to defect, especially after the other agent defected once.

Robustness checks. To make sure that the observed unforgivingness was not due to the particular prompt used, we run several versions of the game as robustness checks, randomizing the order of the presented options, relabelling the choice options and changing the presented utilities to be represented by either points, dollars or coins (Fig. 4). We also repeated our analysis with two different cover stories, added explicit end goals to our prompt, ran games with longer playing horizons and described numerical outcomes with text (also see Supplementary Fig. 3). The results of these simulations showed that the reluctance to forgive was not due to any particular characteristics of the prompts. A crucial question was if GPT-4 did not understand that the other agent wanted to cooperate again or if it could understand the pattern but just did not act accordingly. We, therefore, run another version of the game, where we told GPT-4 explicitly that the other agent would defect once but otherwise cooperate. This resulted in GPT-4 choosing to defect throughout all rounds, thereby maximizing its own points.

Prompting techniques to improve observed behaviour. One problem of these investigations in the Prisoner's Dilemma is that defecting can under specific circumstances be seen as the optimal, utility-maximizing and equilibrium option even in a repeated version, especially if one knows that the other player will always choose to cooperate and when the number of interactions is known. Thus, we run more simulations to assess if there could be a scenario in which GPT-4 starts to forgive and cooperates again, maximizing the joint benefit instead of its own.

We took inspiration from the literature on human forgiveness in the Prisoner's Dilemma and implemented a version of the task in the vein of ref. 11. Specifically, ref. 11 showed that telling participants that other players sometimes make mistakes makes people more likely to forgive and cooperate again after another player's defection (albeit in infinitely played games). Indeed, this can be favourable to them in terms of pay-offs. We observed similar behaviour in GPT-4 as it started cooperating again.

Battle of the Sexes. In our large-scale analysis, we saw that the different LLMs did not perform well in games that required coordination between different players. In humans, it has frequently been found that coordination problems can be solved by the formation of conventions[25,26].

https://doi.org/10.1038/s41562-025-02172-y

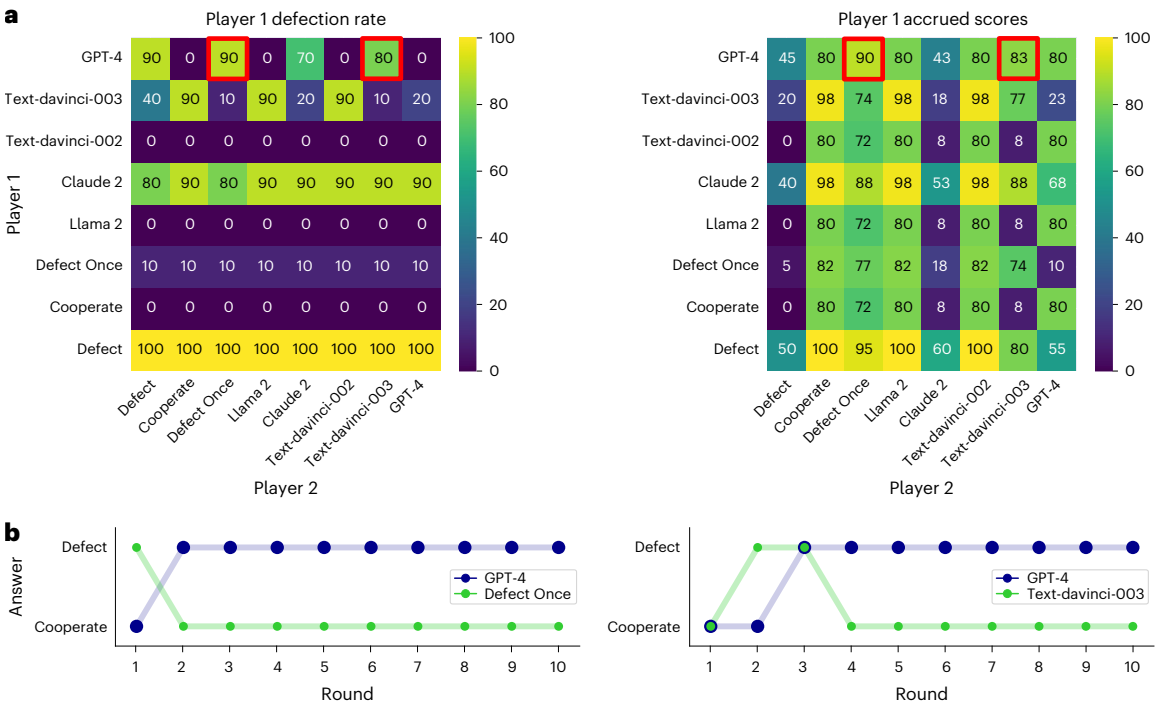

Fig. 3 | Overview of the Prisoner's Dilemma. a, Heat maps showing the player 1 defection rate in each combination of players and the scores accrued by player 1 in each game. **b**, Example gameplays between GPT-4 and an agent that defects once and then cooperates, and between GPT-4 and text-davinci-003. These games are also highlighted in red in the heat maps.

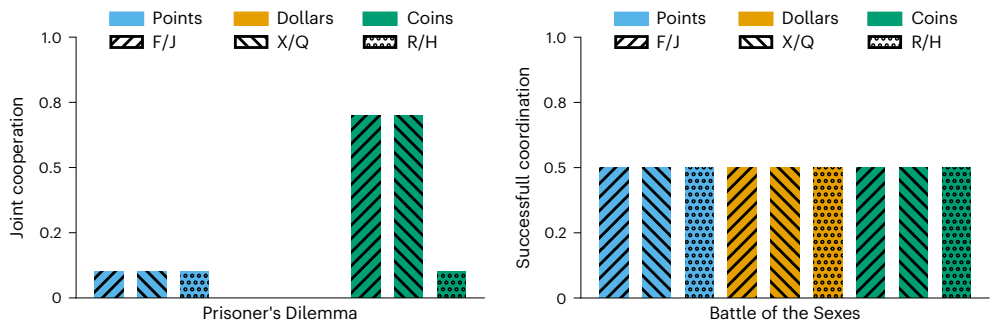

Fig. 4 | Prompt variations. Left: GPT-4's performance for different prompt variations in the Prisoner's Dilemma game against a false defector agent. The probability of joint cooperation is ≤0.1 for all combinations except for two using coins as utility outcomes. Right: GPT-4's performance for different prompt variations in the BoS game against an alternating agent. GPT-4 always chooses its preferred option, resulting in successful coordination rates of only 0.5 across all combinations. For each variation, two random letters that occur with similar frequency in English are given as the choice options.

A coordination game is a type of simultaneous game in which a player will earn a higher pay-off when they select the same course of action as another player. Usually, these games do not contain a pure conflict, that is, completely opposing interests, but may contain slightly diverging rewards. Coordination games can often be solved via multiple pure strategies, or mixed, Nash equilibria in which players choose (randomly) matching strategies. Here, to probe how LLMs balance coordination and self-interest, we look at a coordination game that contains conflicting interests.

We study a game that is archaically referred to as the Battle of the Sexes, a game from the family of biased games. Assume that a couple wants to decide what to do together. Both will increase their utility by spending time together. However, while the wife might prefer to watch a football game, the husband might prefer to go to the ballet. Because the couple wants to spend time together, they will derive no utility by doing an activity separately. If they go to the ballet together, or to a football game, one person will derive some utility by being with the other person but will derive less utility from the activity itself than the other person. The corresponding pay-off matrix is

$$\begin{array}{lcc} & \text{Football} & \text{Ballet} \\ \text{Football} & (10,7) & (0,0) \\ \text{Ballet} & (0,0) & (7,10) \end{array}. \quad (2)$$

As before, the playing agents are all three versions of GPT, Claude 2 and Llama 2 as well as three more simplistic strategies. For the simplistic strategies, we implemented two agents who always choose just one option. Because LLMs most often interact with humans, we additionally implemented a strategy that mirrored a common pattern exhibited by human players in the battle of the sexes. Specifically, humans have been shown to often converge to turn-taking behaviour in the Battle of the Sexes[27–30]; this means that players alternate between jointly picking the better option for one player and picking the option for the other

https://doi.org/10.1038/s41562-025-02172-y

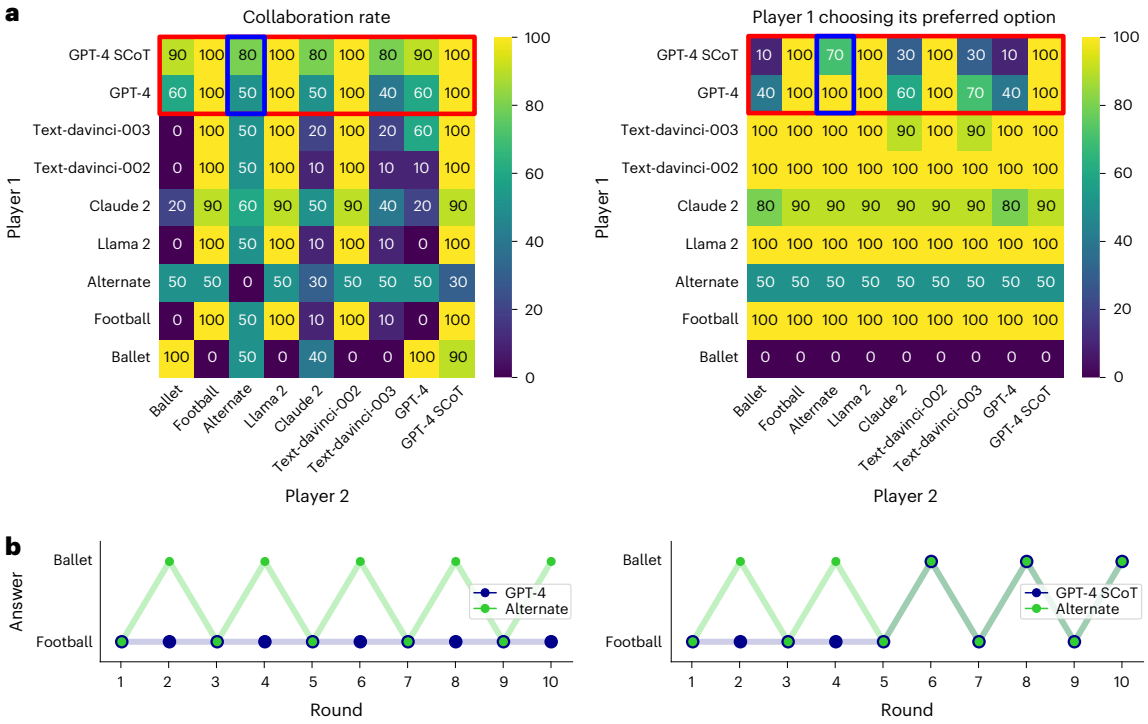

Fig. 5 | Overview of the Battle of the Sexes. a, Heat maps showing rates of successful collaboration between the two players and the rates of player 1 choosing its preferred option football. GPT-4 SCoT and GPT-4 performance comparisons are highlighted in red. **b**, Gameplay between GPT-4 and an agent that alternates between the two options (left) and gameplay between GPT-4 and GPT-4 SCoT that represents a GPT-4 model prompted using the SCoT method to first predict the opponent's move before making its own move by reasoning about its prediction (right). Both games are also highlighted in blue in the heat maps.

player. While not a straightforward equilibrium, this behaviour has been shown to offer an efficient solution to the coordination problem involved and to lead to high joint welfare[28].

Figure 5 shows the results of all interactions. As before, GPT-4 plays generally better than all other agents ($t(128.28) = 2.83$, $P = 0.005$, $d = 0.28$, 95% CI 0.07–0.50, BF 3.56). Yet, while GPT-4 plays well against other agents who choose only one option, such as an agent always choosing football, it does not play well with agents who frequently choose their non-preferred option. For example, when playing against text-davinci-003, which tends to frequently choose its own preferred option, GPT-4 chooses its own preferred option repeatedly but also occasionally gives in and chooses the other option. Crucially, GPT-4 performs poorly when playing with an alternating pattern (where, for courtesy, we let agents start with the option that the other player preferred). This is because GPT-4 seemingly does not adjust its choices to the other player but instead keeps choosing its preferred option. GPT-4, therefore, fails to coordinate with a simple, human-like agent, an instance of a behavioural flaw.

Robustness checks. To make sure that this observed behavioural flaw was not due to the particular prompt used, we also rerun several versions of the game, where we randomize the order of the presented options, relabelled the choice options and changed the presented utilities to be represented by either points, dollars or coins as shown in Fig. 4. We also repeated our analysis with two different cover stories, in which we told GPT-4 that it was taking part in a cooking competition or working on a collaborative project keeping the underlying problem structure (pay-offs and the interaction dynamics) identical (Supplementary Fig. 3). The results of these simulations showed that the inability to alternate was not due to any particular characteristics of the used prompts. To make sure that the observed behavioural flaw was not due to the particular pay-off matrix used, we also rerun several versions of the game, where we modified the pay-off matrix gradually from preferring football to preferring ballet (or, in our case, the abstract F and J). The results of these simulations showed that GPT-4 did not alternate for any of these games but simply changed its constant response to the option that it preferred for any particular game. Thus, the inability to alternate was not due to the particular pay-off matrix we used (Supplementary Section A.5).

Prediction scenarios. Despite these robustness checks, another crucial question remains: Does GPT-4 simply not understand the alternating pattern or can it understand the pattern but is unable to act accordingly? To answer this question, we run two additional simulations. In the first simulation, GPT-4 was again framed as a player in the game itself. However, we now additionally ask it to predict the other player's next move according to previous rounds. In this simulation, GPT-4 started predicting the alternating pattern correctly from round 5 onwards (Fig. 6).

In the second simulation, instead of having GPT-4 be framed as a player itself, we simply prompted it with a game between two ('external') players and asked it to predict one player's next move according to the previous rounds. For the shown history, we used the interaction between GPT-4 and the alternating strategy. In this simulation, GPT-4 started predicting the alternating pattern correctly even earlier, from round 3 onwards. Thus, GPT-4 seemingly could predict the alternating patterns but instead just did not act in accordance with the resulting convention. Similar divergences in abilities between social and non-social representations of the same situation have been observed in autistic children[31].

SCoT prompting. Finally, we wanted to see if GPT-4's ability to predict the other player's choices could be used to improve its own actions. This idea is closely related to how people's reasoning in repeated games

https://doi.org/10.1038/s41562-025-02172-y

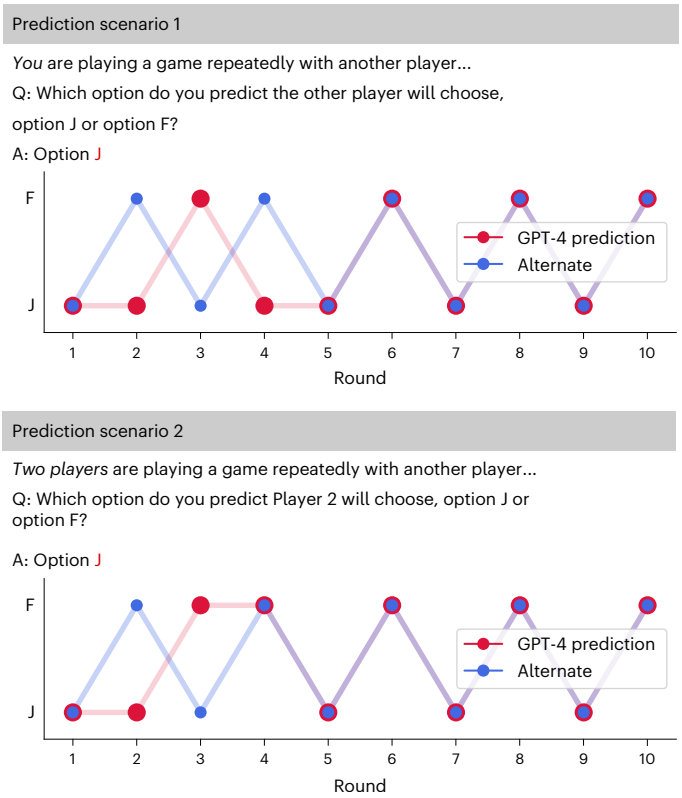

Fig. 6 | Prediction scenarios in the Battle of the Sexes. Top: GPT-4 is a player of the game and predicts the other player's move. Bottom: GPT-4 is a mere observer of a game between player 1 and player 2 and predicts player 2's move.

and tasks about other agents' beliefs can be improved[32]. For example, computer-aided simulations to improve the social reasoning abilities of autistic children normally include questions to imagine different actions and outcomes[33]. This has been successfully used to improve people's decision-making more generally. It is also in line with the general finding that chain-of-thought prompting improves LLM's performance, even in tasks measuring theory of mind (ToM)[34]. Thus, we implemented a version of this reasoning through actions by asking LLMs to imagine the possible actions and their outcomes before making a decision. We termed this approach SCoT prompting. Applying this method improved GPT-4's behaviour, and it started to alternate from round 5 onwards (Fig. 5).

Human experiments

Given the behavioural signatures observed in GPT-4's responses in the different games, we were interested in how actual human subjects would behave when playing with such agents. To test this, we conducted an experiment in which 195 participants played both the Battle of the Sexes and the Prisoner's Dilemma against LLMs. Because the SCoT prompting turned out to be a most reliable modification of LLMs' behaviour, we applied this prompting method only in our behavioural experiments with humans.

Participants were told that they would play either against a human player or an artificial agents for ten repeated rounds for each game and, after each game, had to guess whether they had played against a human or not. Which game they played first was assigned randomly. While all subjects, in fact, played only against LLMs, one group played against the base version of GPT-4, while another group played against a version of GPT-4 that first predicted the other agent's move and the acted accordingly, that is, SCoT prompting. Importantly, each participant played only two games, and the prompting was reset between games to ensure any change in LLM behaviour was not influenced by prior interactions within the experiment. If assigned to the base version initially, participants played both games with this model, and likewise for the socially prompted version. An overview of the experimental design is shown in Fig. 7a. Participants were recruited from Prolific and debriefed fully after the experiment. We were interested in how people played against LLMs in general as well as if GPT-4's behaviour could be improved via SCoT prompting. Finally, we also asked participants whether they thought they had played with another human or an artificial agent after each game.

While participants' average score was significantly higher for the SCoT-prompted condition compared with the condition without further prompting (that is, base) in the Battle of the Sexes (mixed-effects regression results: $\beta = 0.74$, $t(193) = 3.49$, $P < 0.001$, 95% CI 0.32–1.15, BF 80.6), no such difference was observed in the Prisoner's Dilemma ($\beta = 0.10$, $t(193) = 0.47$, $P = 0.64$, 95% CI −0.31 to 0.51, BF 0.2). Looking at the behaviour of both players, we found that SCoT prompting increased successful coordination (that is, both players picking the same option) in the Battle of the Sexes ($\beta = 0.33$, $z = 3.59$, $P < 0.001$, 95% CI 0.15–0.51, BF 13.4), while it also slightly increased joint cooperation (that is, both players cooperating) in the Prisoner's Dilemma ($\beta = 0.24$, $z = 2.54$, $P = 0.01$, 95% CI 0.05–0.42, BF 6.5). In general, participants were more likely to think that the prompted model was another human player as compared with the unprompted base GPT-4 model ($\beta = 0.54$, $z = 8.31$, $P < 0.001$, 95% CI 0.05–0.42, BF 17.6). Additional analysis on participants' temporal behaviour in both games can be found in the Supplementary Information.

In summary, SCoT prompting can increase GPT-4's coordination and cooperation behaviour without changing scores in scenarios where self-interest is important for good behaviour, that is, the Prisoner's Dilemma, but leading to increased performance in coordination problems, that is, the Battle of the Sexes.

Discussion

LLMs are among the most quickly adopted technologies ever, interacting with millions of consumers within weeks[10]. Understanding in a more principled manner how these systems interact with us, and with each other, is thus of urgent concern. Here, our proposal is simple: Just like behavioural game theorists use tightly controlled and theoretically well-understood games to understand human interactions, we use these games to study the interactions of LLMs.

We thereby understand our work as both a proof of concept of the utility of this approach and an examination of the individual failures and successes of socially interacting LLMs. Our large-scale analysis of all 2 × 2 games highlights that the most recent LLMs indeed are able to perform well on a wide range of game-theoretic tasks as measured by their own individual reward, particularly when they do not have to explicitly coordinate with others. This adds to a wide-ranging literature showcasing emergent phenomena in LLMs[4–8]. However, we also show that LLMs' behaviour is suboptimal in coordination games, even when faced with simple strategies.

To tease apart the behavioural signatures of these LLMs, we zoomed in on two of the most canonical games in game theory: the Prisoner's Dilemma and the Battle of the Sexes. In the Prisoner's Dilemma, we show that GPT-4 plays mostly unforgivingly. Starting with full cooperation, it permanently shifts to defection after a single negative interaction with the other agent, even if the other agent later cooperates. While noting that GPT-4's continual defection is indeed the equilibrium policy in this finitely played game, such behaviour comes at the cost of the two agents' joint pay-off. We see a similar tendency in GPT-4's behaviour in the Battle of the Sexes, where it has a strong tendency to stubbornly stick with its own preferred alternative. In contrast to the Prisoner's Dilemma, this behaviour is suboptimal, even on the individual level.

Current generations of LLMs are generally assumed, and trained, to be benevolent assistants to humans[35]. Despite many successes in this

https://doi.org/10.1038/s41562-025-02172-y

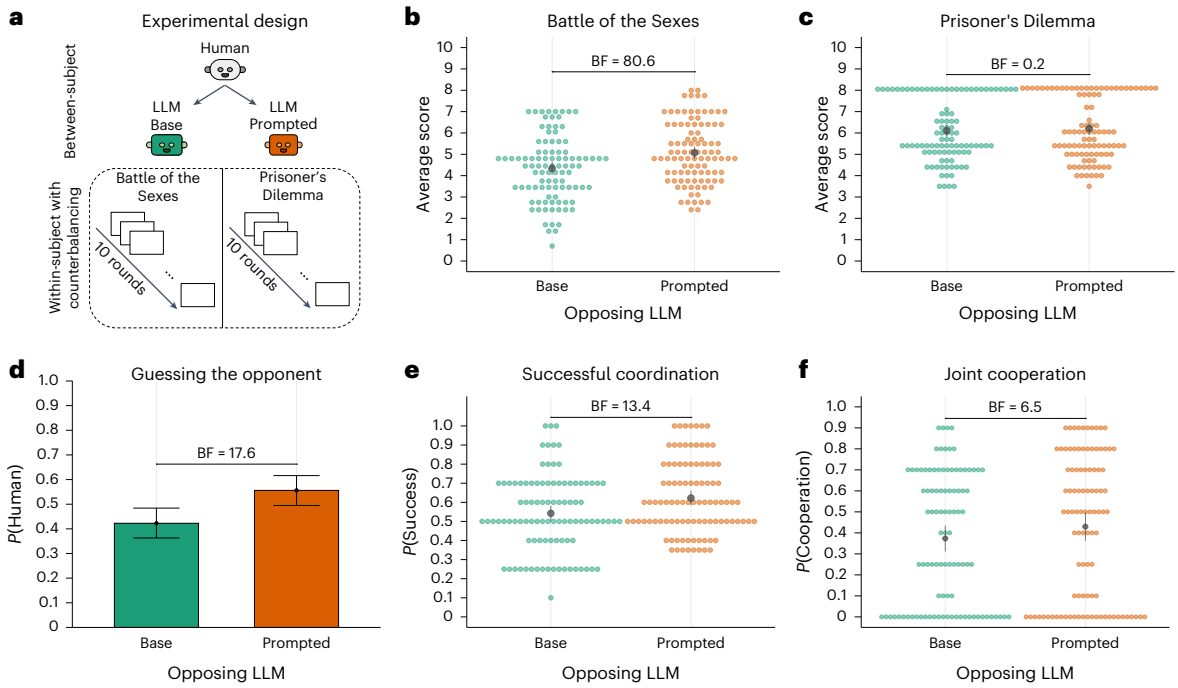

Fig. 7 | Human experiments. a, The design of human experiments ($N = 195$, 89 females, mean age 26.72, s.d. 4.19). Each participant gets randomly assigned either the base or the SCoT-prompted version of the LLM at the start and plays both games repeatedly for ten rounds against this agent. **b**, Results of the Battle of the Sexes game showing participants' average scores by condition (mixed-effects regression results: $\beta = 0.74$, $t(193) = 3.49$, $P < 0.001$, 95% CI 0.32–1.15, BF 80.6). **c**, Results of the Prisoner's Dilemma game showing participants' average scores by condition ($\beta = 0.10$, $t(193) = 0.47$, $P = 0.64$, 95% CIs −0.31 to 0.51, BF 0.2). **d**, The average proportion of participants guessing that they have played against another human by condition. Error bars represent the 95% CIs of the mean ($\beta = 0.54$, $z = 8.31$, $P < 0.001$, 95% CI 0.05–0.42, BF 17.6). **e**, Participants' successful coordination rates by condition in the Battle of the Sexes game ($\beta = 0.33$, $z = 3.59$, $P < 0.001$, 95% CI 0.15–0.51, BF 13.4). **f**, Participants' mutual cooperation rates by condition in the Prisoner's Dilemma game ($\beta = 0.24$, $z = 2.54$, $P = 0.01$, 95% CI 0.05–0.42, BF 6.5).

direction, the fact that we here show how they play iterated games in such a selfish and uncoordinated manner sheds light on the fact that there is still substantial ground to cover for LLMs to become truly social and well-aligned machines[36]. Their lack of appropriate responses vis-a-vis even simple strategies in coordination games also speaks to the recent debate around ToM in LLMs[37–39] by highlighting a potential failure mode.

Our extensive robustness checks demonstrate how these behavioural signatures are not functions of individual prompts but reflect broader patterns of LLM behaviour. Our intervention pointing out the fallibility of the playing partner—which leads to increased cooperation—adds to a literature that points to the malleability of LLM social behaviour in tasks to prompts[40,41]. This is important as we try to understand what makes LLMs better, and more pleasant, interactive partners. Further experiments on GPT-4's final round behaviour have shown that it did not adjust its behaviour in the last round of games or when faced with varying probabilities of continuation, unlike human players who often increase cooperation when future interactions are likely[42,43]. This suggests that GPT-4 may lack mechanisms for backward induction and long-term strategic planning, primarily focusing on immediate context due to its training on next-token prediction[44]. Consequently, GPT-4 tends to default to defection in uncertain situations, contrasting with human tendencies to anticipate and adjust based on future outcomes[24,45].

We additionally observed that prompting GPT-4 to make predictions about the other player before making its own decisions can alleviate behavioural flaws and the oversight of even simple strategies. This represents a more explicit way to force an LLM to engage in ToM and shares much overlap with non-SCoT reasoning[34,46]. Just like chain-of-thought prompting is now implemented as a default in some LLMs to improve (non-social) reasoning performance, our work suggests implementing a similar social cognition prompt to improve human–LLM interaction.

In our exploration of a behavioural game theory of machines, we acknowledge several limitations. First, despite covering many families of games, our investigation is constrained to simple 2 × 2 games. However, we note that our analysis substantially goes beyond current investigations that have often investigated only one game, and done so using single-shot rather than iterated instances of these games. For example, our iterated approach shares more overlap with the more iterated nature of human–LLM conversations. We also note that we mainly study finite games where agents share knowledge about the duration of the interaction. This is in contrast to so-called indefinite games that have either unknown, probabilistic or no endpoints at all. In these games, both optimal prescriptions and empirical behaviour can differ significantly from the finite case, warranting further investigation.

We believe that more complicated games will shed even more light on game-theoretic machine behaviour in the future. For example, games with more continuous choices like the trust game[47] might elucidate how LLMs dynamically develop (mis-)trust. Games with more than two agents, like public goods or tragedy of the commons type games[48], could probe how 'societies' of LLMs behave, and how LLMs cooperate or exploit each other.

Given the social nature of the tasks studied here, further empirical work is needed to fully understand human–LLM interactions across all paradigms. In our study, we conducted human experiments in two of the games, specifically, the Battle of the Sexes and the Prisoner's Dilemma, and attempted to probe human-like behaviours such as turn-taking in Battle of the Sexes or prompting for forgiveness in the Prisoner's Dilemma. However, these empirical investigations were limited to these two games. By extending human studies to the remaining games, additional dynamics may emerge. Furthermore, asking LLMs

to self-report their strategies in these games and correlating these explanations with their actions could provide valuable insights into their actual decision-making processes.

Our results highlight the broader importance of a behavioural science for machines[49–52]. We believe that these methods will continue to be useful for elucidating the many facets of LLM cognition, particularly as these models become more complex, multimodal and embedded in physical systems.

Related work

As algorithms become increasingly more able and their decision making processes impenetrable, the behavioural sciences offer new tools to make inferences just from behavioural observations[49,50]. Behavioural tasks have, therefore, been used in several benchmarks[10,53].

Whether and how algorithms can make inferences about other agents, machines and otherwise, is one stream of research that borrows heavily from the behavioural sciences[54–56]. Of particular interest to the social interactions most LLMs are embedded in is an ability to reason about the beliefs, desires and intentions of other agents, or a ToM[57]. ToM underlies a wide range of interactive phenomena, from benevolent teaching[58] to malevolent deception[56,59], and is thought to be the key to many social phenomena in human interactions[60,61].

Whether LLMs possess a ToM has been debated. For example, it has been argued that GPT-3.5 performs well on a number of canonical ToM tasks[39]. Others have contested this view, arguing that such good performance is merely a function of the specific prompts[37,38]. Yet, other research has shown that chain-of-thought reasoning improves LLMs' ToM ability[34]. Moreover, the currently largest LLM, GPT-4, manages to perform well in ToM tasks, including in the variants in which GPT-3.5 previously struggled[8]. Thus, GPT-4's behaviour will be of particular interest in our experiments.

Games taken from game theory present an ideal testbed to investigate interactive behaviour in a controlled environment[62], and LLMs' behaviour has been probed in such tasks[63]. For example, ref. 40 let GPT-3 participate in the dictator game, and ref. 41 used the same approach for the ultimatum game. Both show how the models' behaviour is malleable to different prompts, for example, making them more or less self-interested. However, all these games rely on single-shot interactions over fewer games and do not use iterated games.

Our study builds upon recent advancements in the field, which have shifted the focus from solely assessing the performance of LLMs to comparing them with human behaviours. Previous research efforts have explored various approaches to analyse LLMs, such as using cognitive psychology tools[51,64] and even adopting a computational psychiatry perspective[52].

Finally, the theory behind interacting agents is important for many machine learning applications in general[65] and, in particular, in adversarial settings[66], where one agent tries to trick the other agent into thinking that a generated output is good. Understanding prosocial dynamics in multiagent systems[67] and fostering cooperation in them[68] is essential for developing robust and trustworthy artificial intelligence systems that can navigate complex social environments[69].

Methods

To investigate how human subjects would behave when playing with LLM agents, we studied their interactions in two of the games we used: Prisoner's Dilemma and the Battle of the Sexes. We also investigated if participants could detect and behave differently when playing against different agents. Participants (N = 195, 89 females, mean age 26.72, s.d. 4.19) were recruited through Prolific[70], an online platform that allows researchers to access a diverse and reliable pool of participants. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications[71–73]. The participants were required to be fluent speakers of English with minimum approval rates of 0.95 and 1, and a minimal number of previous submissions of 10 that have not participated in our experiment before. All participants provided informed consent before inclusion in the study. Experiments were performed in accordance with the relevant guidelines and regulations approved by the ethics committee of the University of Tübingen (protocol no. 701/2020BO). Participants received a £3 base payment plus a bonus of up to £2 depending on performance (1 cent for each point received during the games) for their participation. The average compensation was £11.41 per hour. Participants were fully debriefed after the experiment. Data of 21 players who failed to make a round's choice between the two options within a given time frame (20 s) were excluded.

In the sections that follow, we first detail the experimental set-up for LLM–LLM interactions, which serves as a comparative baseline for our study. We then present details from the human participant study outlined above.

LLM–LLM interactions

We study LLMs' behaviour in finitely repeated games with full information taken from the economics literature. We focus on two-player games with discrete choices between two options to simplify the analyses of emergent behaviours. We let two LLMs interact via prompt chaining, that is, all integration of evidence and learning about past interactions happens as in-context learning[4,74]. The games are submitted to LLMs as prompts in which the respective game, including the choice options, is described. At the same time, we submit the same game as a prompt to another LLM. We obtain generated tokens $\mathbf{t}$ from both LLMs by sampling from

$$p_{\mathrm{LLM}}(\mathbf{t}|\mathbf{c}^{(p)}) = \prod_{k=1}^{K} p_{\mathrm{LLM}}\left(t_k|c_1^{(p)}, \ldots, c_n^{(p)}, t_1, \ldots, t_{k-1}\right). \quad (3)$$

After feeding the prompt to the LLM, our methodology is as follows. The LLM prediction of the first token following the context is $d = p_{\mathrm{LLM}}(t_1|\mathbf{c}^{(p)})$ and the N tokens for the possible answers of the multiple choice question are $o = \{o_i\}_{i=1}^{N}$ which in this case are J and F. The predicted option is then given by

$$\hat{o} = \arg\max(\hat{c}_i), \text{with } \hat{c}_i = d[c_i], i = 1 \ldots N, \quad (4)$$

which are the predicted probabilities of the language model. Once both LLMs have made their choices, which we track as a completion of the given text, we update the prompts with the history of past interactions as concatenated text and then submit the new prompt to both models for the next round. These interactions continue for ten rounds in total for every game. In a single round, $\pi_i(x_1, x_2)$ is the pay-off for player 1 when x_1 and x_2 are the strategies chosen by both players. In repeated games, the pay-offs are often considered as discounted sums of the pay-offs in each game stage, using a discount factor δ. If the game is repeated n times, the pay-off U_i for player i is

$$U_i = \pi_i(x_{10}, x_{20}) + \delta \times \pi_i(x_{11}, x_{21}) + \delta^2 \times \pi_i(x_{12}, x_{22}) + \ldots + \delta^{n-1} \times \pi_i(x_{1(n-1)}, x_{2(n-1)}). \quad (5)$$

Each term represents the discounted pay-off at each stage of the repeated game, from the first game ($t = 0$) to the nth game ($t = n - 1$). In our experiments, we keep $\delta = 1$. To avoid influences of the particular framing of the scenarios, we provide only barebones descriptions of the pay-off matrices (see example in Fig. 1). To avoid contamination through particular choice names or the used framing, we use the neutral options F and J throughout[51].

Games considered. We first investigate 144 different 2 × 2 games where each player has two options, and their individual reward is a function of their joint decision. These games can be categorized into six distinct families—win–win, Prisoner's Dilemma family, unfair, cyclic, biased

and second-best—each with unique characteristics and outcomes. A win–win game is a special case of a non-zero-sum game that produces a mutually beneficial outcome for both players provided that they choose their corresponding best option. They encourage cooperation, leading to outcomes where both parties benefit. In brief, in games from the Prisoner's Dilemma family, two agents can choose to work together, that is, cooperate, for average mutual benefit, or betray each other, that is, defect, for their own benefit. The typical outcome is a Nash equilibrium that is suboptimal for both players compared with a possible Pareto-superior outcome. In an unfair game, one player can always win when playing properly, leading to highly unequal outcomes. Cyclic games are characterized by the absence of dominant strategies and equilibria. In these games, players can cycle through patterns of choices without settling into a stable outcome. Biased games are games where agents get higher points for choosing the same option but where the preferred option differs between the two players. One form of a biased game is the Battle of the Sexes, where players need to coordinate to choose the same option. Finally, second-best games are games where both agents fare better if they jointly choose the option that has the second-best utility. In many of these games, strategic swaps in pay-offs can alter the game dynamics, potentially converting them into different types of game. For two additional games, Prisoner's Dilemma and Battle of the Sexes, we also let LLMs play against simple, hand-coded strategies to understand their behaviour in more detail.

LLMs considered. In this work, we evaluate five LLMs. For all of our tasks, we used the public OpenAI API with the GPT-4, text-davinci-003 and text-davinci-002 models, which are available via the completions endpoint, Meta AI's Llama 2 70B chat model, which has 70 billion parameters and is optimized for dialogue use cases, and the Anthropic API model Claude 2 to run our simulations. Experiments with other popular open-source models MosaicPretrainedTransformer (MPT), Falcon and different versions of Llama 2 (namely MPT-7B, MPT-30B, Falcon-7b, Falcon-40b, Llama 2 7B and Llama 2 13B) have revealed that these models did not perform well at the given tasks, choosing the first presented option more than 95% of the time independent of which option this is. Therefore, we chose not to include them in our main experiments. For all models, we set the temperature parameters to 0 and only ask for one token answer to indicate which option an agent would like to choose. All other parameters are kept as default values.

Playing 6 families of 2 × 2 games task design. While 2 × 2 games can appear simple, they present some of the most powerful ways to probe diverse sets of interactions, from pure competition to mixed motives and cooperation, which can further be classified into canonical subfamilies outlined elegantly by ref. 22. Here, to cover the wide range of possible interactions, we study the behaviours of GPT-4, text-davinci-003, text-davinci-002, Claude 2 and Llama 2 across these canonical families. We let all five engines play all variants of games from within the six families.

Cooperation and coordination task design. We then analyse two games, Prisoner's Dilemma and Battle of the Sexes, in more detail because they represent interesting edge cases where the LLMs performed exceptionally well, and relatively poorly. We hone in particularly on GPT-4's behaviour because of recent debates around its ability for ToM, that is, whether it is able to hold beliefs about other agents' intentions and goals, a crucial ability to successfully navigate repeated interactions[8,39]. For the two additional games, we also let LLMs play against simple, hand-coded strategies to further understand their behaviour. These simple strategies are designed to assess how LLMs behave when playing with more human-like players.

Statistical tests. All reported tests are two-sided. We also report Bayes factors quantifying the likelihood of the data under H_A relative to the likelihood of the data under H_0. We calculate the default two-sided Bayesian t-test using a Jeffreys–Zellner–Siow prior with its scale set to $\sqrt{2}/2$, following[75]. For parametric tests, the data distribution was assumed to be normal, but this was not formally tested. We report effect sizes as either Cohen's d or standardized regression estimates, including their 95% CIs.

Human–LLM interactions

The following sections provide additional details on the design and conduct of the human participant study, including compensation, demographics, prompting and the cover stories.

Design. Experiments were presented to participants using a combination of HTML, JavaScript and CSS with custom code. After a presentation of the instructions including screenshots from the actual gameplay, participants were required to complete a comprehension questionnaire. Only upon responding correctly to all questions could they proceed to the main part of the experiment. Participants played both the Prisoner's Dilemma and the Battle of the Sexes, with the order counter-balanced between subjects. Participants were instructed that they would play two games with ten rounds each with different players. The participants' interface (Supplementary Fig. 4) was designed to provide clear and actionable information about the current game. After each game, participants were asked to indicate if they thought they had just played with another human player or an artificial agent.

Prompts and human instructions. The cover story used for interactions with both LLMs and human participants was content-wise identical, including the rules of the game and the history of previous interactions, to ensure consistent framing across conditions (see Supplementary A.1 for the detailed prompt progression). However, the presentation was adapted to suit each audience. For human participants, visual cues and concise text were prioritized to create a more engaging experience (Supplementary Fig. 4).

Ending and debriefing. Participants were informed that their opponent could either be another human participant or an artificial agent. In reality, all participants were paired with either a SCoT-prompted or an unprompted version of GPT-4 for the entirety of the experiment, that is, across both games. After completing the study, participants were debriefed that the purpose of the study was to explore how to make LLMs more human-like and that, in both games, they had played against different versions of an artificial agent.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All participant and model simulation data from the experiments are available via GitHub at https://github.com/eliaka/repeatedgames.

Code availability

The code underlying this study, prompt variations and model simulations is available via GitHub at https://github.com/eliaka/repeatedgames.

References

1. Brants, T., Popat, A., Xu, P., Och, F. J. & Dean, J. Large language models in machine translation. In *Proc. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* 858–867 (Association for Computational Linguistics, 2007).

2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Vol. 1 (Long and Short Papers)* 4171–4186 (Association for Computational Linguistics, 2019).
3. Radford, A. et al. *Improving Language Understanding by Generative Pre-training* (OpenAI, 2018).
4. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inform. Process. Syst.* **33**, 1877–1901 (2020).
5. Wei, J. et al. Emergent abilities of large language models. Preprint at https://arxiv.org/abs/2206.07682 (2022).
6. Webb, T., Holyoak, K. J. & Lu, H. Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **7**, 1526–1541 (2023).
7. Chen, M. et al. Evaluating large language models trained on code. Preprint at https://arxiv.org/abs/2107.03374 (2021).
8. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at https://arxiv.org/abs/2303.12712 (2023).
9. Coda-Forno, J. et al. Meta-in-context learning in large language models. *Adv. Neural Inform. Process. Syst.* **36**, 65189–65201 (2023).
10. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at https://arxiv.org/abs/2108.07258 (2021).
11. Fudenberg, D., Rand, D. G. & Dreber, A. Slow to anger and fast to forgive: cooperation in an uncertain world. *Am. Econ. Rev.* **102**, 720–749 (2012).
12. Mailath, G. J. & Morris, S. Coordination failure in repeated games with almost-public monitoring. *Cowles Foundation Discussion Papers*, 1761 https://elischolar.library.yale.edu/cowles-discussion-paper-series/1761 (2004).
13. Camerer, C. F. *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton University Press, 2011).
14. Fudenberg, D. & Tirole, J. *Game Theory* (MIT Press, 1991).
15. Von Neumann, J. & Morgenstern, O. *Theory of Games and Economic Behavior* (Princeton Univ. Press, 1944).
16. Camerer, C. F. Progress in behavioral game theory. *J. Econ. Perspect.* **11**, 167–188 (1997).
17. Henrich, J. et al. In search of homo economicus: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* **91**, 73–78 (2001).
18. Rousseau, D. M., Sitkin, S. B., Burt, R. S. & Camerer, C. Not so different after all: a cross-discipline view of trust. *Acad. Manag. Rev.* **23**, 393–404 (1998).
19. Johnson, T. & Obradovich, N. Measuring an artificial intelligence language model's trust in humans using machine incentives. *J. Phys. Complex.* **5**, 015003 (2024).
20. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
21. Owen, G. *Game Theory* (Emerald Group Publishing, 2013).
22. Robinson, D. & Goforth, D. *The Topology of the 2 × 2 Games: A New Periodic Table* vol. 3 (Psychology Press, 2005).
23. Jones, G. Are smarter groups more cooperative? Evidence from prisoner's dilemma experiments, 1959–2003. *J. Econ. Behav. Org.* **68**, 489–497 (2008).
24. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
25. Hawkins, R. X. & Goldstone, R. L. The formation of social conventions in real-time environments. *PLoS ONE* **11**, e0151670 (2016).
26. Young, H. P. The economics of convention. *J. Econ. Perspect.* **10**, 105–122 (1996).
27. Andalman, A. & Kemp, C. *Alternation in the Repeated Battle of the Sexes*. 9.29, Spring 2004, MIT (MIT Press, 2004).
28. Lau, S.-H. P. & Mui, V.-L. Using turn taking to mitigate coordination and conflict problems in the repeated battle of the sexes game. *Theory Decis.* **65**, 153–183 (2008).
29. McKelvey, R. D. & Palfrey, T. R. *Playing in the Dark: Information, Learning, and Coordination in Repeated Games* (California Institute of Technology, 2001).
30. Arifovic, J. & Ledyard, J. Learning to alternate. *Exp. Econ.* **21**, 692–721 (2018).
31. Swettenham, J. What's inside someone's head? Conceiving of the mind as a camera helps children with autism acquire an alternative to a theory of mind. *Cogn. Neuropsychiatry* **1**, 73–88 (1996).
32. Westby, C. & Robinson, L. A developmental perspective for promoting theory of mind. *Top. Lang. Disord.* **34**, 362–382 (2014).
33. Begeer, S. et al. Theory of mind training in children with autism: a randomized controlled trial. *J. Autism Dev. Disord.* **41**, 997–1006 (2011).
34. Moghaddam, S. R. & Honey, C. J. Boosting theory-of-mind performance in large language models via prompting. Preprint at https://arxiv.org/abs/2304.11490 (2023).
35. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inform. Process. Syst.* **35**, 27730–27744 (2022).
36. Wolf, Y., Wies, N., Levine, Y. & Shashua, A. Fundamental limitations of alignment in large language models. Preprint at https://arxiv.org/abs/2304.11082 (2023).
37. Ullman, T. Large language models fail on trivial alterations to theory-of-mind tasks. Preprint at https://arxiv.org/abs/2302.08399 (2023).
38. Le, M., Boureau, Y.-L. & Nickel, M. Revisiting the evaluation of theory of mind through question answering. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 5872–5877 (Association for Computational Linguistics, 2019).
39. Kosinski, M. Evaluating large language models in theory of mind tasks. *Proc. Natl Acad. Sci. USA* **121**, e2405460121 (2024).
40. Horton, J. J. *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* (National Bureau of Economic Research, 2023).
41. Aher, G. V., Arriaga, R. I. & Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning* 337–371 (PMLR, 2023).
42. Dal Bó, P. & Fréchette, G. R. The evolution of cooperation in infinitely repeated games: experimental evidence. *Am. Econ. Rev.* **101**, 411–429 (2011).
43. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).
44. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (2019).
45. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
46. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inform. Process. Syst.* **35**, 24824–24837 (2022).
47. Engle-Warnick, J. & Slonim, R. L. The evolution of strategies in a repeated trust game. *J. Econ. Behav. Org.* **55**, 553–573 (2004).
48. Rankin, D. J., Bargum, K. & Kokko, H. The tragedy of the commons in evolutionary biology. *Trends Ecol. Evol.* **22**, 643–651 (2007).
49. Rahwan, I. et al. in *Machine Learning and the City: Applications in Architecture and Urban Design* 143–166 (John Wiley & Sons, 2022).
50. Schulz, E. & Dayan, P. Computational psychiatry for computers. *iScience* **23**, 101772 (2020).
51. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl Acad. Sci. USA* **120**, e2218523120 (2023).

52. Coda-Forno, J. et al. Inducing anxiety in large language models increases exploration and bias. Preprint at https://arxiv.org/abs/2304.11111 (2023).
53. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural Inform. Process. Syst.* **35**, 22199–22213 (2022).
54. Rabinowitz, N. et al. Machine theory of mind. In *Proc. 35th International Conference on Machine Learning* (eds Dy, J. & Krause, A.) 80:4218–4227 (PMLR, 2018).
55. Cuzzolin, F., Morelli, A., Cirstea, B. & Sahakian, B. J. Knowing me, knowing you: theory of mind in AI. *Psychol. Med.* **50**, 1057–1061 (2020).
56. Alon, N., Schulz, L., Dayan, P. & Rosenschein, J. A (dis-)information theory of revealed and unrevealed preferences. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems* https://openreview.net/pdf?id=vcpQW_fGaj5 (2022).
57. Frith, C. & Frith, U. Theory of mind. *Curr. Biol.* **15**, R644–R645 (2005).
58. Vélez, N. & Gweon, H. Learning from other minds: an optimistic critique of reinforcement learning models of social learning. *Curr. Opin. Behav. Sci.* **38**, 110–115 (2021).
59. Lissek, S. et al. Cooperation and deception recruit different subsets of the theory-of-mind network. *PLoS ONE* **3**, e2023 (2008).
60. Hula, A., Montague, P. R. & Dayan, P. Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS Comput. Biol.* **11**, e1004254 (2015).
61. Ho, M. K., Saxe, R. & Cushman, F. Planning with theory of mind. *Trends Cogn. Sci.* **26**, 959–971 (2022).
62. Han, T. A., Perret, C. & Powers, S. T. When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games. *Cogn. Syst. Res.* **68**, 111–124 (2021).
63. Chan, A., Riché, M. & Clifton, J. Towards the scalable evaluation of cooperativeness in language models. Preprint at https://arxiv.org/abs/2303.13360 (2023).
64. Lampinen, A. K. et al. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus* **3**, pgae233 (2024).
65. Crandall, J. W. & Goodrich, M. A. Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning. *Mach. Learn.* **82**, 281–314 (2011).
66. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
67. Santos, F. P. Prosocial dynamics in multiagent systems. *AI Mag.* **45**, 131–138 (2024).
68. Guo, H. et al. Facilitating cooperation in human-agent hybrid populations through autonomous agents. *iScience* **26**, 108179 (2023).
69. Powers, S. T. et al. The stuff we swim in: regulation alone will not lead to justifiable trust in AI. *IEEE Technol. Soc. Mag.* **42**, 95–106 (2023).
70. Palan, S. & Schitter, C. Prolific.ac—a subject pool for online experiments. *J. Behav. Exp. Finance* **17**, 22–27 (2018).
71. Normann, H.-T. & Wallace, B. The impact of the termination rule on cooperation in a prisoner's dilemma experiment. *Int. J. Game Theory* **41**, 707–718 (2012).
72. Charness, G. & Rabin, M. Understanding social preferences with simple tests. *Q. J. Econ.* **117**, 817–869 (2002).
73. Wong, R. Y.-m & Hong, Y.-y Dynamic influences of culture on cooperation in the prisoner's dilemma. *Psychol. Sci.* **16**, 429–434 (2005).
74. Liu, P. et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55**, 1–35 (2023).
75. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16**, 225–237 (2009).

Acknowledgements

This work was supported by grants from the Max Planck Society (E.A., L.S., J.C.-F. and E.S.), the Volkswagen Foundation (E.S.), the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2064/1 (grant no. 390727645, E.A., S.J.O. and M.B). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting E.A.

Author contributions

E.A., L.S. and E.S. conceived experiments. E.A. conducted the experiments. E.A. and E.S. analysed the results with input from L.S., J.C.-F. and M.B. E.A., L.S. and E.S. wrote the manuscript with input from S.J.O. and M.B. All authors reviewed the manuscript.

Funding

Open access funding provided by Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02172-y.

Correspondence and requests for materials should be addressed to Elif Akata.

Peer review information *Nature Human Behaviour* thanks Elias Fernández Domingos and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Elif Akata

Last updated by author(s): Feb 25, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For our model tasks, we used the public OpenAI API with the GPT-4, text-davinci-003 and text-davinci-002 models which are available via the completion endpoint, Meta AI's Llama 2 70B Chat model, and the Anthropic API model Claude 2 to run our simulations. Human data was collected via Prolific using an online experiment written in JavaScript/HTML/CSS.
Data analysis	Behavioural data was analysed using R (4.4.0) and Python (3.11). The model simulations were implemented in Python. The code underlying this study, prompt variations and model simulations are available on github.com/eliaka/repeatedgames.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All participant and model simulation data from the experiments are publicly available on GitHub (https://github.com/eliaka/repeatedgames).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Sex and gender were not relevant in study design. The study was distributed evenly to male and female participants. Sex and gender was determined based on participant demographics on the Prolific platform.
Population characteristics	N=195, 89 females, mean age=26.72, SD=4.19.
Recruitment	Participants were recruited from Prolific and were required to be fluent speakers of English. The study may exhibit self-selection bias since participants voluntarily joined via Prolific and were limited to fluent English speakers, potentially reducing generalisability to non-English speaking populations. However, the behavioral games used are based on universal decision-making principles. The use of English speakers was due to language proficiency, not a theoretical limitation.
Ethics oversight	Ethics committee of the Eberhard Karls University Tübingen (protocol nr. 701/2080BO)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative study. We let large language models play finitely repeated games with each other, with human-like strategies and human players using two-player, two-action games from behavioural economics.
Research sample	Our sample consisted of participants (N=195, mean age=26.72, SD=4.19.) recruited via Prolific, with gender counterbalanced, all required to be fluent in English. While not fully representative of the general population due to self-selection and language criteria, this sample was chosen to ensure clear communication and reliable engagement with the behavioral game theory tasks. Additionally, the study featured five language models: OpenAI's GPT-4, text-davinci-003, text-davinci-002, Meta AI's Llama 2 70B Chat, and Anthropic's Claude 2. All models played 2x2 games against each other and hand-coded human-like strategies, with GPT-4 acting as the LLM opponent in the human participant study.
Sampling strategy	Participants were recruited via Prolific using a convenience sampling approach, with stratification to ensure gender counterbalancing. The sample size of N=195 was chosen based on comparable behavioral game theory studies in the literature. Participants played both the Prisoner's Dilemma and the Battle of the Sexes, with the order counterbalanced between subjects. Models were chosen so that they adequately represent current SOTA models (large and small, open source and closed source).
Data collection	The open-source models were evaluated on a Slurm-based cluster with a single A100. For proprietary models, we used the public APIs. Human participant data was collected on Prolific. Data collection code is available on GitHub (github.com/eliaka/repeatedgames). The researcher was not blinded to experimental conditions. In the model experiments, the language models played the games autonomously. In the human study, the design did not require reciprocal interaction with the researcher and it compared two model conditions (prompted vs. baseline). All participants were fully debriefed after the experiment.
Timing	The experiments on Prolific took place between 17/06/2024 and 21/06/2024.
Data exclusions	We excluded data of 21 players who failed to make a round's choice between the 2 options within a given time frame (20 seconds).
Non-participation	14 participants did not complete the study; two of them having internet connection issues, one quoting technical problems and the rest without explicit reasoning.
Randomization	Participants were randomly allocated into groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging