Can Large Language Models Infer Human Actions and Motives in Strategic Decision Making?

Kaleen Shrestha, Harish Dukkipati, Abhinav Gupta, Zhonghao Shi, Maja Matarić

Viterbi School of Engineering
University of Southern California

Abstract

Humans are able to predict each other's actions by reasoning about the others' underlying goals, preferences, and motives, such as greed and risk-aversion. Game theory provides a framework for studying human behaviors through incentivized games that simulate social situations. While prior work has used game theory to explore the game-playing capabilities and behaviors of large language models (LLMs), there has been no systematic investigation of how LLMs infer human actions and motives in games, and whether their reasoning process is similar to that of humans. To address this gap, we utilized two validated games from the cognitive science literature—the Social Prediction Game (SPG) and the Inspection Game (IG)—to systematically study how well several recent open- and closed-source LLMs predict player actions and whether they can leverage and generalize the players' motives learned from the iterated games. Our results indicate that state-of-the-art LLMs can achieve accuracy close to human levels in predicting players' actions with underlying human motives in SPG. However, unlike humans, who rely on reasoning about players' motives to inform their predictions, LLMs appear to depend on zero-shot prediction and fail to recognize the statistical patterns in players' actions. As a result, its prediction accuracy does not improve over multiple rounds. Our results in IG further demonstrate that, unlike humans, LLMs are unable to recognize a player's underlying motives and generalize their understanding of the same player to a new context. This suggests that LLMs lack the reasoning capabilities associated with theory of mind (ToM). Our findings offer insights into differences in human and LLM reasoning mechanisms; in particular, LLMs demonstrate a lack of ToM reasoning, suggesting that further research into human-AI alignment is needed before utilizing LLMs for human behavior modeling and simulation in this and related contexts.

Introduction

People naturally infer others' mental models and predict their future actions in social or strategic situations (Thornton, Weaverdyck, and Tamir 2019). Modeling how humans create mental models of others has been a long-standing research area in cognitive science and artificial intelligence (AI) (Tenenbaum et al. 2011). Past work has utilized game theory to hypothesize the mechanisms humans use to infer others' underlying preferences, or motives, from past gameplay and to subsequently successfully predict the future gameplay actions of others (van Baar et al. 2022). As AI models such as large language models (LLMs) continue to improve in natural language understanding and social reasoning performance, the AI community has been increasingly interested in studying and comparing LLM and human behavior in games (Fontana, Pierri, and Aiello 2024; Fan et al. 2024; Xie et al. 2024). Past works have investigated LLMs behavior in games, but not how large language models (LLM) predict human actions in the context of strategic decision-making.

To predict future actions of others, one makes predictions based on generalizing from past interactions, but each round of interaction may bring in new information that requires adaptation (Tenenbaum 1998). Studies in cognitive science have shown that people approach this generalization vs. adaptation dilemma by inferring the other person's latent underlying motives, such as greed and risk-aversion, that generalizes across different interaction game settings to drive the decision making process (Poncela-Casasnovas et al. 2016).

To study how well LLM agents can predict strategic human actions, we leverage an experimental framework from the cognitive science literature, known as the Social Prediction Game (SPG). Introduced by van Baar et al. (2022), this framework has been used to study how people predict others' future actions in strategic settings. Our findings indicate that LLM models appear to rely on assumptions and biases about human behavior to predict actions of a particular simulated player, similar to humans. However, they fail to reason effectively or to model strategic decision-making processes based on past actions, unlike humans.

Furthermore, in order to investigate the generalizability of an LLM's "mental model" of human motives, we leverage another existing game framework, the Inspection Game (IG) (Avenhaus, Von Stengel, and Zamir 2002). Following the experimental setup introduced by van Baar et al. (2022), IG is used to study the ability for people to generalize their mental model of a player's behavior in a new context. Despite similar accuracy to humans in SPG, we find that LLMs are not able to generalize underlying motives in IG, indicating that LLMs may not be learning information about latent motives of the player in SPG.

In summary, we investigate the following:

RQ 1: Can LLMs predict actions of a simulated player in a strategic decision game?

RQ 2: Can LLMs generalize learned latent motives of simulated player to new contexts?

To answer these research questions, we compare five recent closed-source and six open-source LLMs with different model architectures, training data, training techniques, and model sizes. We additionally contribute LLM prompting techniques for SPG and IG, and plan to release the codebase for this project as part of the final paper.

Background

We utilized two rigorous experimental frameworks utilized in the work by van Baar et al. (2022) to study the learning mechanisms used by humans to predict the actions of simulated players in economic games. The Social Prediction Game (SPG) framework consists of four canonical economic games (Prisoner's Dilemma, Stag Hunt, Harmony Game, and Snowdrift Game) to study humans' ability to predict actions of a simulated player. The Inspection Game (IG) framework (Avenhaus, Von Stengel, and Zamir 2002) focuses on applications related to surveillance under resource-constrained contexts, to study generalization of learned motives from SPG in a new context.

Social Prediction Game (SPG)

SPG consists of 16 rounds of a single-shot game with two players: Player A and an Opponent. In each round, the players choose one of two actions: `COOPERATE` or `DEFECT`. Player A's choices are modeled using a particular set of predefined rule-based motives based on human data (Poncela-Casasnovas et al. 2016) and artificial motives not observed in human data. Further information of these motives is found in the next section. The Opponent's action is randomly selected from the two available options and is not the focus of the game. The LLM serves as an observer of the game and is prompted to predict Player A's action for each round based on the previous round. Player A's behavior is not dependent on the Opponent's actions, but is based on its own underlying rule-based motive.

For each round, one of the four classic economic games is selected randomly as the game for that round. Player A is presented with the set of payoffs for all combinations of possible actions, in text format (Figure 1. An example of the prompt can be found in the Appendix. Player A's decision for that particular set of payoffs is pre-determined based on its motive for that particular SPG and can be visualized by the matrix in Figure 2. The four games are characterized by the relationship between the reward (R) for both players cooperating, penalty (P) for both players defecting, points for the cooperating player (S), and points for the defecting player (T). Following the work of van Baar et al. (2022), we set $R = 10$, and $P = 5$. The four games are defined by the different values for S and T. Each game induces different trade-offs between cooperation and competition.

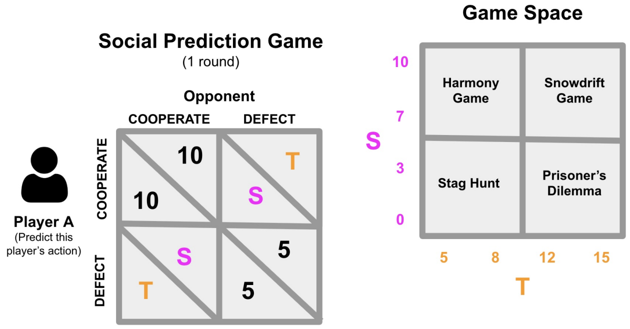


Figure 1: Definition of SPG (left) and the game space (right). The payoff matrices resulting from the different values of S and T lead to four distinct games.

Motives

The goal of the LLM is to predict the action of the simulated Player A in a round in the SPG. The motive selected for Player A is grounded in cognitive science research of human behavior in economic games. Human choice data show that when playing economic games, people tend to follow distinct behaviors that optimize for certain underlying motives (Poncela-Casasnovas et al. 2016). Therefore, following van Baar et al. (2022), we simulated four motives for Player A: greedy, risk-averse, inverse greedy, and inverse risk-averse. As shown in Figure 2, the motives are defined for each of the four economic games, making Player A's actions deterministic based on the chosen motive for a given SPG.

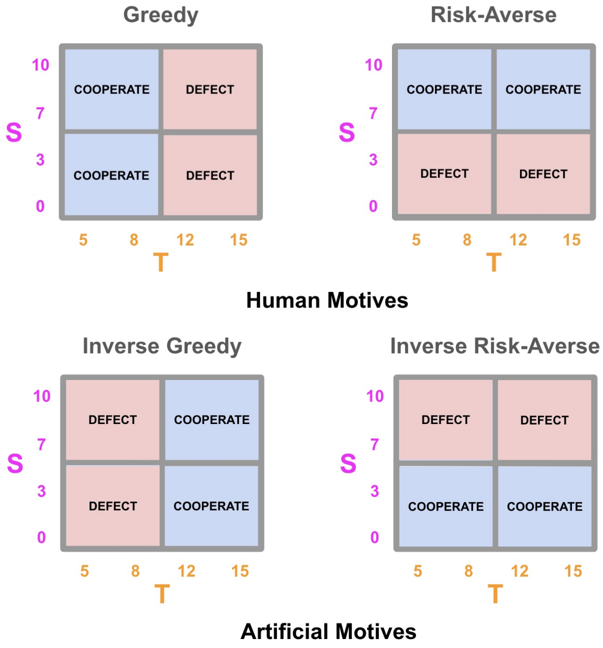


Figure 2: Definitions of player motives based on actions for the four economic games in SPG. Blue cells denote `COOPERATE`, and red cells denote `DEFECT`. These are the human (above) and artificial motive definitions (below).

Poncela-Casasnovas et al. (2016) found that humans can exhibit behaviors associated with greedy and risk-averse motives in economic games. We refer to these as "human motives". Inverse greedy and inverse risk-averse are inverses of the human motives and are not found in human behavior data, so we refer to them as "artificial motives". They are included to add complexity to the social prediction task and cross check whether LLMs pick up on human cognitive and social biases.

Inspection Game (IG)

van Baar et al. (2022) investigated how people learn the latent motive of Player A in SPG by assessing if people can generalize the player's motive to a new context: the Inspection Game (IG). They find that as SPG accuracy increases, so does the generalization of the motive to IG. In this work, we utilize IG as a way to test if the LLM is the learning underlying human motive in the SPG and can generalize this to a new context.

In the experimental setup for IG, the LLM first completes SPG (for a participant with either latent greedy or risk-averse motive) where the LLM predicts Player A's actions. Then, the LLM plays IG *with* Player A. To compare generalization of risk-averse and greedy motives, the LLM then completes SPG for the other motive with a new player, Player B, and completes IG with Player B.

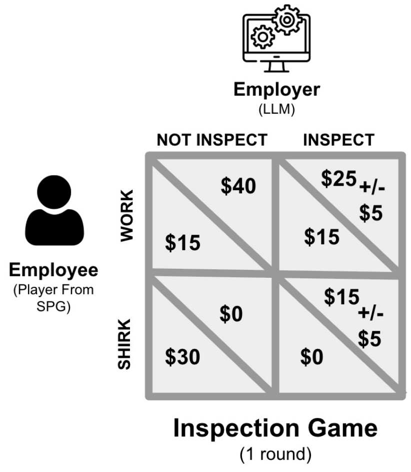


Figure 3: Definition of IG. The LLM is participating in the game as the Employer, choosing to `INSPECT` or `NOT INSPECT` the Employee (the player from the SPG played before the IG). The player is either risk-averse and always chooses `WORK`, or greedy and always chooses `SHIRK`. The goal is to see if the LLM chooses to inspect according to the underlying motive of the Employee learned during SPG. Following a staircase design, the cost to inspect goes up and down by $5 depending on whether the LLM chooses `INSPECT` (reward to inspect goes down) or `NOT INSPECT` (reward to inspect goes up) in order to identify the point of indifference (i.e., find the threshold at which the Employer is willing to pay to inspect).

LLM plays IG against the participant that the LLM was predicting actions for from the previous SPG (Figure 3). The LLM is the Employer while the participant is the Employee. The Employee can either choose `WORK` or `SHIRK`, while the Employer (LLM) can choose `INSPECT` or `NOT INSPECT`. As seen in the payoff matrix (this time, values are dollars instead of points) in Figure 3, there is a greater reward for the Employee if the Employee chooses to shirk and the Employee chooses to not inspect. For the Employer, it is costly to inspect, and therefore there is only a benefit if they inspect when the Employee shirks. The idea is that a greedy player tries to maximize their maximum payoff, and so they always choose to shirk. Conversely, a risk-averse player tries to maximize their minimum payoff, and so they always choose to work.

IG, like SPG, is also a one-shot game where, based on a payoff matrix, the players choose an action simultaneously. However, the LLM participates and chooses an action based on its belief of what the participant chooses. If the LLM infers the latent motive in SPG, then it should choose to inspect more often for the greedy player, and not inspect as often for the risk-averse player. There are 16 rounds of the IG, with no feedback after each round of what the Employee actually chose. For each round, the cost to inspect for the Employer increases by $5 if the Employer chose to inspect, and decreased by $5 if the Employer chose not inspect, in order to find the threshold at which the Employer is willing to pay to inspect (i.e., staircase procedure used by van Baar et al. (2022)). This provides an estimate of strength in the belief of the underlying motive (willing to pay a higher amount to inspect means the Employer is more confident that the Employee greedy and thus shirking). The LLM is also asked directly to output the probability that the Employee is working based on the interactions observed in SPG.

Methodology

To answer the two research questions, we evaluate multiple recent open- and closed- source LLMs on SPG (RQ 1)and IG (RQ 2). We follow the evaluation methods used by van Baar et al. (2022). For SPG, we compare the accuracy of LLMs for predicting players with human and artificial motives. For IG, we compare the probabilities that the player is working output by the LLM for risk-averse and greedy players. The amount of money that the LLM is willing to pay to inspect the Employee is calculated by averaging the cost for inspection in the last 5 rounds (as done by van Baar et al. (2022)) of IG (i.e., determining the indifference point). We compare the LLMs with the human data from van Baar et al. (2022). Additionally, as a baseline for both human and LLM behavior on SPG, we train a purely statistical machine learning model, a random forest classifier, on SPG by representing SPG as a binary classification problem. The 16-round set up was recreated by training the classifier on the $1, 2, ..., ith$ round of data to predict the label for the $i+1th$ round. When comparing differences in outcome measures between two motives for a particular LLM, we used either the two-tailed paired-samples t-test if the data were normally distributed, or Wilcoxon signed-rank test if the data are not normally distributed. When checking for monotonicity of accuracy over all rounds in SPG, we used either Pearson correlation coefficient (r) or Spearman's rank correlation coefficient (ρ) to test for monotonicity. All experiments with LLMs and the random forest classifier were run for five iterations to take into account computational resource constraints while also being able to account for any variance in results.

Large Language Models (LLMs)

We use several recent open- and closed- source models for our experiments, following recent LLM evaluation papers related to game theory (Xie et al. 2024; Fan et al. 2024; Fontana, Pierri, and Aiello 2024) as well as more recently released models. We perform experiments with current state of the art proprietary models from OpenAI (GPT-4o Hurst et al. (2024a) and GPT-4.1 Achiam et al. (2023)), Anthropic (Claude Sonnet 4 Anthropic (2025)), and Google (Gemini 2.5 Comanici et al. (2025)). For comparison, we also tested four smaller, open-source model families: LlaMA 3.3 70B/4 Maverick (Meta 2025a,b), DeepSeek Llama 70B (Guo et al. 2025), Mistral Saba 24B (Team 2025), and OLMo 2 7B/13B(OLMo et al. 2024). These models were selected to represent LLMs with a variety of model architectures (mixture of experts, autoregressive models) and training techniques and data (reinforcement learning from human feedback, chain of thought reasoning, instruction data). We also selected a variety of model sizes, ranging from three billion parameter models to 70 billion parameter models (with possibly up to trillions of parameters for the proprietary models). Input context window size also varies from 4,096 to one million tokens. The models were either downloaded from HuggingFace (Wolf et al. 2019) or used via the cloud service Groq (Groq 2025). We used the default temperature of 1 for all LLMs in order to test zero-shot learning.

Prompts

We added system-level prompts for the rules in the SPG and IG, and user-level prompts for the round-level information. The full prompts we use can be found in the Appendix.

We used a text completion style prompt to mimic the masked language learning objective typically used to pre-train language models. Occasionally models produced additional prompt text; prompts were manually screened and re-formatted to the preset prediction labels: `COOPERATE`/`DEFECT` and `INSPECT`/`NOT INSPECT`.

We also conduct experiments with a subset of models where the previous LLM answers were not included, and found that models performed better when their past guesses were added to the input context. The action Player A (with a latent motive) selected was input into the context history for the next round. The Appendix provides examples of multiple rounds in the prompt for SPG and IG, as well as additional experiments with various prompt and SPG game settings..

RQ 1: Can LLMs Predict Actions of a Simulated Player in a Strategic Decision Game?

In this research question, we contrast LLM accuracy on SPG to human accuracy, and investigate whether LLM accuracy on SPG improves as the rounds progress.

Difference Between Predicting Actions for Human and Artificial Motives

There is a significant gap in accuracy (two-tailed paired-samples t-test $t(149) = 22.0, p < 0.001$ (van Baar et al. 2022)) for SPG between predicting actions for players with human motives (72%) vs. players with artificial motives (47%). As van Baar et al. (2022) also note, people appear to predict players' actions by inferring their rational strategic motives (i.e., the human motives). When players are modeled with counter-intuitive or irrational motives (i.e., the artificial motives), human players struggle to reason about their intentions, and their accuracy tends to be close to random, as expected. A purely statistical learning approach using a random forest classifier, however, only has a 1% accuracy difference between players with artificial and human motives and predicts both with nearly perfect accuracy. This is expected since the motives are step functions that can be modeled by a statistical machine learning model (Figure 2).

We see that all closed-source models except Gemini 2.5 Flash have higher accuracies for SPG with human motives than for SPG with artificial motives, which is similar to the human results, with a statistically significant difference between average accuracy for human motives and artificial motives (GPT-4.1: two-tailed paired-samples t-test $t(4) = -7.12, p = .002$; GPT-4.1 mini: two-tailed paired-samples t-test $t(4) = -5.68, p = .004$; GPT-4o: two-tailed paired-samples t-test $t(4) = -23.73, p < .001$; Claude Sonnet 4: two-tailed paired-samples t-test $t(4) = -11.22, p < .001$; Gemini 2.5 Flash: Wilcoxon signed-rank test $W = 0.0, p = .06$), indicating that these LLMs have leveraged a human bias to predict a player's actions in SPG. Note that the gap, compared to humans, is much larger for GPT-4o, Claude Sonnet 4, and Gemini 2.5 Flash, with the average accuracy for artificial motives being much lower than for human motives. This is an interesting result, as closed-source LLMs have demonstrated similar accuracy to humans on SPG with human motives, but much lower accuracy than humans on the artificial motives, suggesting that the models may be overfitting on human biases. This could be the reason behind the much lower accuracy for SPG with players with artificial motives. Although GPT-4o outperformed the newer GPT-4.1 on human motives, GPT-4.1 and GPT-4.1 mini both have a smaller gap in accuracy between players with human and artificial motives, indicating that the newer GPT models may be slightly better at picking up patterns rather than amplifying human biases.

Most open-source models we tested (Mistral, Deepseek, Llama, OLMo), except the newest Llama 4 Maverick model, perform lower on SPG for human motives compared to people. However, most models, except for OLMo 2 7B Instruct, demonstrated a similar gap in accuracy between human and artificial motives as humans; Llama 3.3 70B (two-tailed paired-samples t-test $t(4) = -5.73, p = .005$), Llama 4 Maverick ($t(4) = -9.60, p < .001$), and OLMo 2 13B ($t(4) = -4.35, p = .01$) showed significant differences between human and artificial motive SPG accuracy. Llama 4 Maverick, the newest Llama model and the only open-source model with a mixture of experts architecture, performed the

Model Family	Model	Social Prediction Game Performance (Acc. % ± Std)	
		Human Motives	Artificial Motives
Statistical	Random Forest Classifier	98 ± 1	97 ± 1
Human (van Baar et al., 2022)	N/A	72 ± 11	47 ± 11
GPT	GPT-4o	73 ± 4	33 ± 5
	GPT-4.1	68 ± 8	41 ± 7
	GPT-4.1 mini	66 ± 14	36 ± 7
Claude	Claude Sonnet 4	71 ± 10	32 ± 5
Gemini	Gemini 2.5 Flash	**74 ± 3**	27 ± 4
Mistral	Mistral-Saba-24B	60 ± 10	**48 ± 5**
Deepseek	Deepseek-llama-70B	53 ± 11	44 ± 11
Llama	Llama 3.3 70B	59 ± 7	36 ± 12
	Llama 4 Maverick	73 ± 12	35 ± 16
OLMo	OLMo 2 7B Instruct	48 ± 9	**48 ± 13**
	OLMo 2 13B Instruct	59 ± 13	39 ± 10

Table 1: Accuracy results of open- and closed- source LLMs compared to human and statistical baselines on SPG. Both humans and most LLMs perform higher on SPG for players with human motives compared to players with artificial motives. Random forest classifier performs similarly on SPG for human and artificial motives.

best on SPG for players with human motives among the rest of the open-source models.

Insight: LLMs we tested perform similarly to humans on SPG, predicting actions of players with human motives with higher accuracy than those with artificial motives.

Humans predict the actions of players with human motives with higher accuracy compared to the actions of those with artificial motives because of social biases. LLMs appear to amplify these social biases in SPG as well. This social bias appears to impair the ability of both humans and LLMs to accurately infer the underlying motive, which is implemented in this context as a step function. Statistical models, such as random forest classifiers, on the other hand, do not leverage any social bias and perform better on SPG for both players, with human and artificial motives. Humans develop stereotypes and generalizations to be more efficient in judging social behaviors, however in cases where there is noise in an underlying motive, this can be a disadvantage. Future development of LLMs for modeling human behavior should take this into account.

Temporal SPG Accuracy

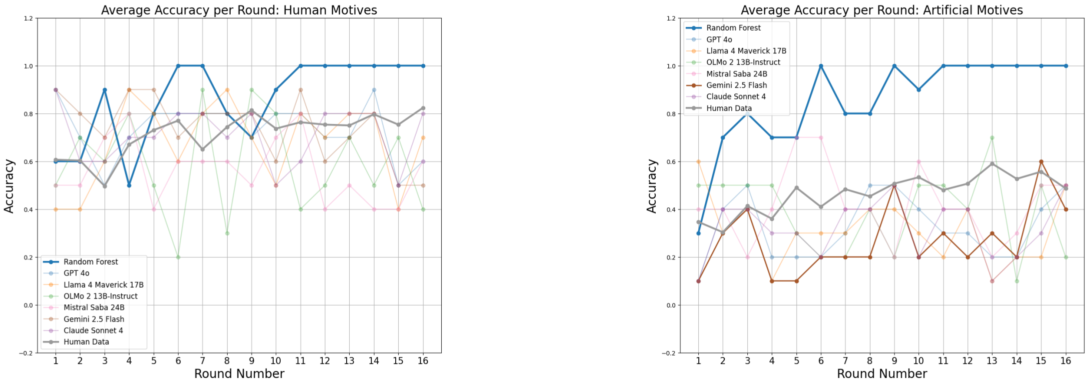
We also investigated how accuracy may improve as the rounds progress in SPG. Humans are good at learning patterns, and with more chances to interact and observe, people usually improve at detecting patterns (Tenenbaum 1998). We explored if LLMs demonstrate this property as well by calculating accuracy at every round of the 16-round SPG.

Figure 4 plots the temporal accuracy for the models from each model family with the highest average accuracy on SPG for human motives. The statistical and human baselines follow an increasing monotonic function as the rounds progress, demonstrating that average accuracy improves as more examples of past player actions are shown for SPG for both human and artificial motives: human motives (human: $\rho(16) = 0.78, p < .001$, random forest: $\rho(16) = 0.74, p < .001$) and artificial motives (human: $r(16) = 0.83, p < .001$, random forest: $\rho(16) = 0.84, p < .001$). However, we see that all LLMs, even the highest performing accuracy models such as GPT-4o and Gemini 2.5 Flash, fail to follow an increasing monotonic function in accuracy for SPG for human (GPT-4o: $r(16) = -0.2, p = 0.45$, Llama 4 Maverick: $\rho(16) = 0.15, p = 0.57$, OLMo 2 13B: $r(16) = -0.07, p = 0.79$, Mistral Saba 24B: $r(16) = -0.24, p = 0.38$, Gemini 2.5 Flash: $r(16) = -0.63, p = 0.01$, Claude Sonnet 4: $\rho(16) = 0.01, p = 0.96$) and artificial motives (GPT-4o: $r(16) = 0.22, p = 0.41$, Llama 4 Maverick: $r(16) = -0.25, p = 0.36$, OLMo 2 13B: $\rho(16) = -0.21, p = 0.43$, Mistral Saba 24B: $r(16) = -0.05, p = 0.86$, Gemini 2.5 Flash: $r(16) = 0.46, p = 0.07$, Claude Sonnet 4: $r(16) = 0.1, p = 0.7$).In fact, we see that Gemini 2.5 Flash accuracy follows a statistically significant *decreasing* monotonic function.

Insight: LLMs we tested do not improve in accuracy in the Social Prediction Game with more observations of the player's actions, indicating a lack of both statistical and ToM reasoning capabilities.

RQ 2: Can LLMs Generalize Learned Latent Motives of Simulated Player to New Contexts?

We investigate LLM performance on IG using only closed-source models due to computational constraints. Risk-averse players always chose `WORK` in IG, while greedy players always chose `SHIRK`, and the goal is to check if the model uses what it observed about the player's motive during the SPG to strategically inspect or not inspect the player during IG. van Baar et al. (2022) found that humans demonstrated this capability to generalize the underlying motive to IG (with statistical significance), which improved with higher accuracy on the SPG, indicating that people are able to generalize a learned underlying motive to a new context. Table 2 shows that most LLMs were able to correctly rank the risk-



(a)

(b)

Figure 4: (a) Accuracy (averaged over different iterations) of the SPG for each round for players with human motives. (b) Accuracy (averaged over different iterations) of the SPG for each round for players with artificial motives. For both motives, the human (gray) and statistical model (blue) accuracy followed a roughly increasing monotonic function as the rounds progressed. LLMs do not follow a monotonically increasing accuracy function.

Model	**Probability % Player is Working (Avg. Prob. ± Std)**		**Willingness to Pay $ to Inspect (Avg. Amount ± Std)**	
	Greedy	**Risk-Averse**	**Greedy**	**Risk-Averse**
Human (van Baar et al. 2022)	41	55	20 ± 9	15 ± 9
GPT-4o	41 ± 5	44 ± 15	5 ± 5	4 ± 5
GPT-4.1	46 ± 13	53 ± 19	4 ± 7	4 ± 7
GPT-4.1 mini	40 ± 6	44 ± 10	7 ± 3	7 ± 3
Claude Sonnet 4	68 ± 6	70 ± 5	0 ± 0	0 ± 0
Gemini 2.5 Flash	54 ± 11	62 ± 20	13 ± 10	13 ± 10

Table 2: Results of five recent closed-sourced models on IG compared to the human baseline results by van Baar et al. (2022). The human results were averaged over 150 participants, the model results were averaged over 5 repetitions of IG. Most models correctly ranked the risk-averse players as more likely to work, however most models conversely did not pay more to inspect greedy players, with the exception of GPT-4o (but the difference was insignificant; two-tailed Wilcoxon signed-rank test: $W = 0.0, P > 0.05$). The values for willingness to pay for the LLM experiments differed from the van Baar et al. (2022) results since we used a different variable amount for the staircase procedure used in IG.

averse player as more likely to work compared to the greedy player, however the difference in rankings was not statistically significant (GPT-4o: two-tailed paired-samples t-test $t(4) = 0.43, p = .69$; GPT-4.1: two-tailed paired-samples t-test $t(4) = 1.34, p = .25$; GPT-4.1 mini: two-tailed paired-samples t-test $t(4) = 0.61, p = .58$; Claude Sonnet 4: two-tailed paired-samples t-test $t(4) = 0.40, p = .71$; Gemini 2.5 Flash: Wilcoxon signed-rank test $W = 1.0, p = 1.0$). Similarly, although the rankings align somewhat with the underlying motive of the player, the amount of money is not reflective of the underlying motive (LLMs should be willing to pay more money to inspect greedy players compared to risk-averse players), with no models showing statistically significant differences in the amount they are willing to pay to inspect for risk-averse versus greedy players (GPT-4o: Wilcoxon signed-rank test $W = 3.0, p = 1.0$; GPT-4.1: two-tailed paired-samples t-test $t(4) = W = 3.0, p = .31$; GPT-4.1 mini: two-tailed paired-samples t-test $t(4) = 0.606478, p = .58$; Claude Sonnet 4: two-tailed paired-samples t-test $t(4) = 0.40, p = .71$; Gemini 2.5 Flash: two-tailed paired-samples t-test $t(4) = 0.86, p = .44$). GPT-4o shows slightly higher willingness to pay for greedy players, however this was not statistically significantly higher than for risk-averse players.

Insight: LLMs we tested were not able to extend learned human motives to a new context in the Inspection Game, indicating that LLMs we tested do not identify the underlying motives in the Social Prediction Game.

These results suggest that LLMs are not able to generalize the underlying motive to new contexts, which puts into question whether LLMs can observe and, in turn, leverage the underlying motive to perform well on SPG. The increasingly long context history may be contributing to the low generalization capabilities, with LLMs possibly forgetting past information observed.

Related Work

LLMs for Modeling Human Behavior

LLMs are increasingly being used to model and simulate human behavior across various contexts. This growing body of research focuses on measuring alignment between LLM outputs and human behavioral patterns, with particular attention to the risks of oversimplifying or misrepresenting the complexity of human decision-making processes.

Recent work has established various methodologies for evaluating how well LLMs capture human behavioral patterns through direct comparison frameworks. Nie et al. (2023) introduced MoCa, a comprehensive framework for measuring human-language model alignment on causal and moral judgment tasks, revealing significant variations in alignment quality across different moral scenarios and demographic groups. Unlike MoCa, which examined static moral judgment tasks, our work focuses on dynamic, sequential prediction where LLMs must infer the underlying motive from sequential gameplay to predict future actions.

Aher, Arriaga, and Kalai (2023) evaluated LLMs as proxies for human participants in behavioral studies, finding some promising correlations but also systematic biases, especially in risk assessment and emotional decisions. In contrast, our work focuses on LLMs' ability to model human strategic decision-making in validated games, predicting actions based on inferred motives rather than replacing human subjects.

LLMs in Game Theory

Game theory provides a mathematical framework for modeling strategic interactions among rational agents. Traditionally, it has been used to study human behavior in simplified, abstract scenarios known as economic games. These games have played a central role in experimental economics and social science research for decades, revealing fundamental behavioral patterns in cooperation and competition (Poncela-Casasnovas et al. 2016). Recently, the emergence of large language models (LLMs) such as GPT-4o (Hurst et al. 2024b) has opened new avenues for investigating the game-theoretic behavior of artificial agents.

Recent work has examined LLMs' ability to model other agents' beliefs, goals, and decisions. ToMBench tested theory of mind (ToM) capabilities of LLMs via scenario-based questions, finding that while models like GPT-4 handled basic false beliefs and intentions, they struggled with second-order reasoning and non-literal cues (Chen et al. 2024). Fan et al. (2024) showed that LLMs often deviate from rational choice theory in games like rock-paper-scissors, revealing issues with belief updating and payoff optimization. TMGBench further found that while LLMs perform well in simple 2x2 games, they underperformed in tasks requiring long-term planning or multi-agent coordination (Wang et al. 2024). Together, these studies suggest LLMs show emergent but limited strategic reasoning.

These works primarily focused on LLM game-playing capabilities, however there has not yet been a systematic study of their ability to infer human actions and investigating the mechanisms for learning to predict human actions. These are fundamental skills for strategic decision-making, and an important step towards understanding LLMs' capabilities and limitations in real-world human-AI interaction. To this end, we focus on investigating this capability of LLMs to model human decision making.

Conclusion

We explore two research questions regarding how LLMs compare to humans in determining underlying motives in SPG and IG. We gain insight on how LLMs appear to amplify human biases by predicting actions of players with human motives with higher accuracy compared to those of players with artificial motives. Furthermore, our results demonstrate that LLMs do not improve in accuracy as the rounds in the SPG progress, which is unlike humans and a standard statistical model. Secondly, we find that LLMs are not able to generalize the learned motive in SPG to a new context in IG, unlike humans. This work identifies differences in the behavior of LLMs when modeling human behavior in the context of game theory. These insights highlight the need for further research into human-AI alignment.

Limitations and Future Work

The SPG and IG prompting methods lack mechanisms to verify whether models understand the game, unlike Fontana, Pierri, and Aiello (2025), who assessed rule comprehension and data parsing in an iterated prisoner's dilemma. In our experiments, while proprietary models adhered to the prompts, some open-source models (e.g., OLMo 2 7B) produced overly verbose outputs. Although we manually reviewed these outputs, a more robust approach would systematically assess game understanding to filter out unsuitable models.

Extending the previous point, a more direct way to check for generalization capabilities of LLMs in IG is to prompt the LLM directly to explain the underlying motive, to check if the underlying motive was identified. Additionally, a more rigorous experimental setup for checking generalization of LLMs in IG would provide the LLM with the underlying motive of the player from SPG to check if the actions chosen in IG align with the motive, and compare to a setting where the underlying motive is not provided.

Additionally, the models we used have variable context window sizes, and so as the number of rounds increases, it is possible that the context exceeded certain models' limits (i.e., the OLMo models with context window sizes of 4,096 tokens). This resulted in truncated inputs and thus prevented models from improving accuracy as the games progressed.

The players in the SPG and IG were simulated using motives from human data (Poncela-Casasnovas et al. 2016), however human motives are not always consistent and free from noise. A more realistic simulation of human behavior or engaging human players would provide more reliable data.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.;

Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aher, G.; Arriaga, R. I.; and Kalai, A. T. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *arXiv preprint arXiv:2208.10264*.

Anthropic. 2025. Claude Sonnet 4. Technical report, Anthropic.

Avenhaus, R.; Von Stengel, B.; and Zamir, S. 2002. Inspection games. *Handbook of game theory with economic applications*, 3: 1947–1987.

Chen, Z.; Wu, J.; Zhou, J.; Wen, B.; Bi, G.; Jiang, G.; Cao, Y.; Hu, M.; Lai, Y.; Xiong, Z.; et al. 2024. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Fan, C.; Chen, J.; Jin, Y.; and He, H. 2024. Can large language models serve as rational players in game theory? a systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17960–17967.

Fontana, N.; Pierri, F.; and Aiello, L. M. 2024. Nicer Than Humans: How do Large Language Models Behave in the Prisoner's Dilemma? *arXiv preprint arXiv:2406.13605*.

Fontana, N.; Pierri, F.; and Aiello, L. M. 2025. Nicer Than Humans: How Do Large Language Models Behave in the Prisoner's Dilemma? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 522–535.

Groq. 2025. Groq Cloud. Technical report, Groq.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024a. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024b. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Meta. 2025a. Llama 3.3. Technical report, Meta.

Meta. 2025b. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Technical report, Meta.

Nie, A.; Zhang, Y.; Amdekar, A.; Piech, C.; Hashimoto, T.; and Gerstenberg, T. 2023. MOCA: Measuring human-language model alignment on causal and moral judgment tasks.

OLMo, T.; Walsh, P.; Soldaini, L.; Groeneveld, D.; Lo, K.; Arora, S.; Bhagia, A.; Gu, Y.; Huang, S.; Jordan, M.; et al. 2024. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*.

Poncela-Casasnovas, J.; Gutiérrez-Roig, M.; Gracia-Lázaro, C.; Vicens, J.; Gómez-Gardeñes, J.; Perelló, J.; Moreno, Y.; Duch, J.; and Sánchez, A. 2016. Humans display a reduced set of consistent behavioral phenotypes in dyadic games. *Science advances*, 2(8): e1600451.

Team, M. A. 2025. Mistral Saba. Technical report, Mistral AI.

Tenenbaum, J. 1998. Bayesian modeling of human concept learning. *Advances in neural information processing systems*, 11.

Tenenbaum, J. B.; Kemp, C.; Griffiths, T. L.; and Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022): 1279–1285.

Thornton, M. A.; Weaverdyck, M. E.; and Tamir, D. I. 2019. The social brain automatically predicts others' future mental states. *Journal of Neuroscience*, 39(1): 140–148.

van Baar, J. M.; Nassar, M. R.; Deng, W.; and Feldman-Hall, O. 2022. Latent motives guide structure learning during adaptive social choice. *Nature Human Behaviour*, 6(3): 404–414.

Wang, H.; Feng, X.; Li, L.; Guo, Y.; Qin, Z.; Sui, D.; and Kong, L. 2024. Tmgbench: A systematic game benchmark for evaluating strategic reasoning abilities of llms. *arXiv preprint arXiv:2410.10479*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xie, C.; Chen, C.; Jia, F.; Ye, Z.; Lai, S.; Shu, K.; Gu, J.; Bibi, A.; Hu, Z.; Jurgens, D.; et al. 2024. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems*, 37: 15674–15729.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the “Type your response here” text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the “Type your response here” text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the \question command or any other lines in this document.

You can \input this .tex file right before \end{document} of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) NA

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) Yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) Yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) No

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) Yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) yes

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes

LLM Experiments

For all LLMs in both SPG and IG, we used default hyperparameters with a temperature of 1 in order ot test zero-shot capabilities of the LLMs on SPG and IG.

0.1 GPT-4o: temperature = 1.0, top-p=1.0

0.2 GPT-4.1: temperature = 1.0, top-p=1.0

0.3 GPT-4.1 mini: temperature = 1.0, top-p=1.0

0.4 Claude Sonnet 4: temperature = 1.0, top-p=0.99

0.5 Gemini 2.5 Flash: temperature = 1.0, top-p=0.95

0.6 Mistral Saba 24B: temperature = 1.0, top-p=0.9

0.7 Deepseek Llama 70B: temperature = 1.0, top-p=0.9

0.8 Llama 3.3 70B: temperature = 1.0, top-p=0.9

0.9 Llama 4 Maverick: temperature = 1.0, top-p=0.9

0.10 OLMo 2 7B Instruct: temperature = 1.0, top-p=0.9

0.11 OLMo 2 13B Instruct: temperature = 1.0, top-p=0.9

To run SPG and IG, we used the following model cards and computing resources:

0.1 GPT-4o: gpt-4o-2024-08-06 - OpenAI API/cloud compute

0.2 GPT-4.1: gpt-4.1-2025-04-14 - OpenAI API/cloud compute

0.3 GPT-4.1 mini: gpt-4.1-mini-2025-04-14 - OpenAI API/cloud compute

0.4 Claude Sonnet 4: claude-sonnet-4-20250514 (no extended thinking) - Anthropic API/cloud compute

0.5 Gemini 2.5 Flash: gemini-2.5-flash - Google Cloud Platform (GCP) API/cloud compute

0.6 Mistral Saba 24B: mistral-saba-24b - Huggingface/ 2 RTX 3090 GPUs with 24GB each

0.7 Deepseek Llama 70B: deepseek-r1-distill-llama-70b - Groq/cloud compute

0.8 Llama 3.3 70B: llama3-70b-8192 - Groq/cloud compute

0.9 Llama 4 Maverick: meta-llama/llama-4-maverick-17b-128e-instruct - Groq/cloud compute

0.10 OLMo 2 7B Instruct: allenai/OLMo-2-1124-7B-Instruct - Huggingface/ 2 RTX 3090 GPUs with 24GB each

0.11 OLMo 2 13B Instruct: allenai/OLMo-2-1124-13B-Instruct - Huggingface/ 2 RTX 3090 GPUs with 24GB each

LLM Prompts for SPG and IG

SPG LLM Prompt Snippet

# System prompt

`<<SYS>>`

Context: Player A is playing a multi-round game against random opponents. At each turn Player A and the opponent simultaneously perform one of the following actions: [COOPERATE, DEFECT].

For each round, Player A will have a completely new opponent. You will be predicting what action A will choose in each round of this game. You will see the history of actions A actually made. Output must be one of the following actions: [COOPERATE, DEFECT].

`<</SYS>>`

# Round-level prompt
In round 1, the payoffs for each combination of chosen actions are the following:

If A plays DEFECT and opponent plays DEFECT, A collects **5** points and opponent collects **5** points.
If A plays DEFECT and opponent plays COOPERATE, A collects **8** points and opponent collects **3** points.
If A plays COOPERATE and opponent plays DEFECT, A collects **3** points and opponent collects **8** points.
If A plays COOPERATE and opponent plays COOPERATE, A collects **10** points and opponent collects **10** points.

You think A plays:

Figure 5: LLM prompt snippet for SPG. There is a system-level prompt, as well as one round-level prompt for the first round that describes the payoff matrix for a particular round from one of the four game types described in Figure 1.

IG LLM Prompt Snippet

# **System prompt**

<<SYS>>

Context: You will now be playing a game against the same Player A from the previous games. In this game, you are the employer and Player A is your employee. There will be multiple rounds of this game, where you will be playing against Player A in all rounds.

Player A will decide whether to shirk or do their work. You will simultaneously decide to either inspect of not inspect whether or not Player A is working, by choosing one of the following actions: [INSPECT, NOT INSPECT].

There will be certain payoffs for the different combinations of actions you and Player A will choose. Consider the payoffs and your belief of whether Player A will work or shirk.

Given the game settings, you will be choosing one of the following actions: [INSPECT, NOT INSPECT].

<</SYS>>

# **Round-level prompt**

In round 1, the payoffs for each combination of chosen actions are the following:
If A chooses WORKS and you choose INSPECT, A gets **15** dollars and you get **25** dollars.
If A chooses WORKS and you choose NOT INSPECT, A gets **15** dollars and you get **40** dollars.
If A chooses SHIRK and you choose INSPECT, A collects **0** dollars and you get **15** dollars.
If A chooses SHIRK and you choose NOT INSPECT, A collects **30** dollars and you get **0** dollars.
You choose:

Figure 6: LLM prompt snippet for IG. There is a system-level prompt, as well as one round-level prompt for the first round that describes the payoff matrix for a particular round based on the staircase procedure outlined in Figure 3.