Can Large Language Models Infer Human Actions and Motives? Evaluation in Social Prediction and Inspection Games

Kaleen Shrestha*, **Abhinav Gupta, Harish Dukkipati, Zhonghao Shi, Maja Matarić**
Department of Computer Science, University of Southern California
***Corresponding author:** kshresth@usc.edu

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. 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 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 SPGs. However, unlike humans, who rely on reasoning about players' motives to inform their predictions, LLMs failed to recognize statistical patterns in players' actions. As a result, LLM prediction accuracy did not improve over multiple rounds. Our results in the IG further demonstrate that, unlike humans, LLMs were unable to recognize a player's underlying motives and to generalize their understanding of the same player to a new context. This suggests that LLMs may lack reasoning capabilities. Our findings offer insights into differences in human and LLM reasoning mechanisms, suggesting that further research into human-AI alignment is needed before utilizing LLMs for human behavior modeling and simulation in this and related contexts.

1 Introduction

People naturally infer others' mental models and predict their future actions in social and strategic situations (Thornton et al., 2019). Modeling how humans create mental models of others is a long-standing research area in cognitive science and AI (Tenenbaum et al., 2011). Past work has utilized game theory to hypothesize the mechanisms humans use to infer others' underlying preferences and motives from past gameplay and to successfully predict 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, the AI community has been increasingly interested in studying and comparing LLM and human behavior in games (Fontana et al., 2024; Fan et al., 2024; Xie et al., 2024). Past works have investigated LLM behavior in games, however there has not been much research into how well LLMs predict human actions in the context of strategic decision-making.

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

To study how well LLMs can predict strategic human actions, we leveraged 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 LLMs display similar limitations in predicting motives as humans do. However, LLMs failed to model strategic decision making processes based on past actions, unlike humans.

Furthermore, to investigate the generalizability of an LLM's "mental model" of human motives, we leveraged another existing game framework, the Inspection Game (IG) (Avenhaus et al., 2002). Following the experimental setup introduced by van Baar et al. (2022), the IG was used to study the

Findings of the Association for Computational Linguistics: ACL 2026, pages 32791–32803

ability to generalize a mental model of a player's behavior in a new context. Despite having similar accuracy to humans in SPGs, we found that LLMs were not able to generalize underlying models of motives in the IG, indicating that LLMs may not be inferring information about latent motives of the player.

In summary, we investigated the following research questions:

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

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

To answer these questions, we compared recent closed-source and open-source LLMs with different model architectures, training techniques, and model sizes. We additionally contribute LLM prompting techniques for SPG and IG, and release our codebase with simulation code, data, and evaluation, for SPG and IG [1].

2 Background

We utilized two rigorous experimental frameworks from the work by van Baar et al. (2022) which studied human learning mechanisms used 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) used to study human players' ability to predict actions of a simulated opponent. The *Inspection Game* (IG) framework (Avenhaus et al., 2002) focuses on applications related to surveillance under resource-constrained contexts, to study generalization of learned motives from SPGs to new contexts.

2.1 Social Prediction Games (SPGs)

SPGs 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.

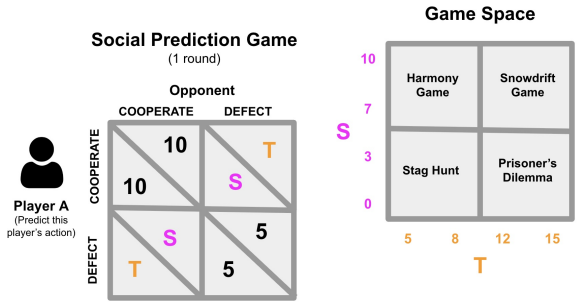


Figure 1: Definition of SPG (left) and the game space (right). On the left, we see the definition of one round of SPG, which is a single-shot game between the Player A of interest, and a random Opponent. The matrix illustrates the points assigned to the players based on the four different action combinations. On the right are the different point values of S and T, which yield four distinct games with different action incentives.

In each round of an SPG, one of the four classic economic games is randomly selected. Player A is presented with the set of payoffs for all combinations of possible actions in that game, 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 predetermined based on their motive for that particular SPG and can be visualized by the matrix in Figure 2. The four SPG 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). Each game induces different trade-offs between cooperation and competition. Following the work of van Baar et al. (2022), we set $R = 10$, and $P = 5$.

2.2 Motives

In each round of an SPG, the goal of the LLM is to predict the action of the simulated Player A. 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 specific underlying motives (Poncela-Casasnovas et al., 2016). Therefore, following van Baar et al. (2022), we simulated four motives for Player A: greedy, risk-

[1] https://github.com/interaction-lab/social-decision-modeling

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.

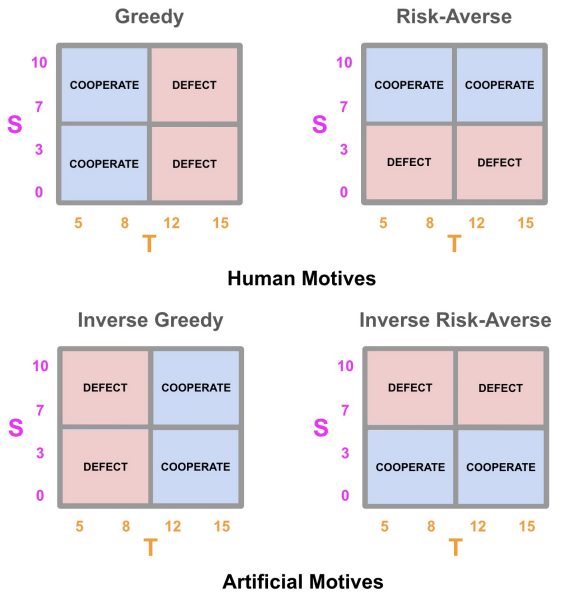


Figure 2: Definitions of human (top) and artificial (bottom) player motives, respectively, based on actions for the four economic games in SPG (see game space matrix on the right in Figure 1): Harmony Game (upper left), Snowdrift Game (upper right), Stag Hunt (lower left), and Prisoner's Dilemma (lower right). Blue cells denote that the player chooses to COOPERATE in that particular game, and red cells denote when they choose to DEFECT. T and S are variables with four different point assignments that define the different games, as shown on the x and y axes of the matrices.

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 the inverses of those two 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 for comparison.

2.3 The Inspection Game (IG)

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

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

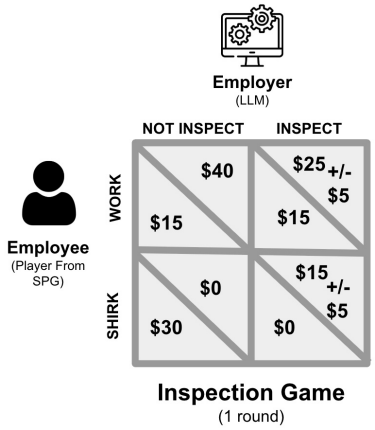


Figure 3: Definition of the IG. The LLM plays the Employer, choosing to INSPECT or NOT INSPECT the Employee (the player from the SPG played before the IG). The player is either risk-averse and always chooses WORK, or greedy and always chooses SHIRK. The goal is to see if the LLM chooses to inspect according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

LLM plays the IG against the player for whom the LLM was predicting actions in 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 (values are dollars) in Figure 3, the reward is greater for the Employee if the Employee chooses to shirk and the Employer chooses For the Employer, it is costly to inspect, and therefore there is only a benefit if inspecting catches the Employee shirking. A greedy Employee tries to maximize the maximum payoff, always choosing to shirk. Conversely, a risk-averse Employee tries to maximize the minimum payoff,

always choosing to work.

The IG, like SPGs, is also a one-shot game where, based on a payoff matrix, the two players choose an action simultaneously. In the LLM version, if the LLM inferred the latent motive in the SPG, then it should choose to inspect more often for a greedy Employee, and not inspect as often for the risk-averse Employee. There are 16 rounds of the IG, with no feedback after each round of what the Employee 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 to inspect, in order to find the threshold at which the Employer is willing to pay to inspect (i.e., a staircase procedure used by van Baar et al. (2022)). This provides an estimate of the strength in the belief of the underlying motive; i.e., willing to pay a higher amount to inspect means the Employer (LLM) is more confident that the Employee is 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 the SPG.

3 Methodology

To answer the two research questions, we evaluated multiple recent open- and closed- source LLMs on SPGs (RQ 1) and the IG (RQ 2). We followed the evaluation methods used by van Baar et al. (2022). For SPGs, we compared LLM accuracy in predicting players with human and artificial motives. For the IG, we compared the probabilities that the player is working provided by the LLM for risk-averse and greedy players. As was done by van Baar et al. (2022), the amount of money that the LLM was willing to pay to inspect the Employee was calculated by averaging the cost for inspection in the last 5 rounds of the IG (i.e., determining the indifference point). We compared LLM performance with human -performance data from van Baar et al. (2022). Additionally, as a baseline for both human and LLM behavior on SPGs, we trained a purely statistical machine learning model, a random forest classifier, on SPGs by representing SPGs as a binary classification problems. 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 were not normally distributed. For all models in this work, we ran 34 repetitions of SPGs and the IG (with a random seed of 27) to achieve sufficient statistical power and determine a medium effect size. When checking for monotonicity of accuracy over rounds in SPGs, we used the Mann-Kendall Test, a nonparametric monotonic trend test for sequential data. We additionally investigated model confidence and used Mann-Kendall Test to test positive monotonicity of model confidence, as well as Pearson correlation and Spearman's rank correlation coefficient to evaluate the relationship between accuracy and model confidence.

3.1 Large Language Models (LLMs) Tested

We tested several recent open- and closed- source models, following recent LLM evaluation work related to game theory (Xie et al., 2024; Fan et al., 2024; Fontana et al., 2024), as well as more recently released models. We conducted experiments with current state-of-the-art proprietary models from Anthropic (Claude Sonnet 4.5 (Anthropic, 2025)), Google (Gemini 2.5 (Comanici et al., 2025)), and OpenAI (GPT 4o (Hurst et al., 2024) and GPT-4.1 (Achiam et al., 2023)). For comparison, we also tested two smaller, open-source models from Meta: Llama 3.3 70B and Llama 4 Maverick 17B with 128 experts (Meta, 2025a,b). These models were selected to represent LLMs with a variety of model architectures (e.g. mixture of experts). We also selected a variety of model sizes, ranging from 17 billion to 70 billion parameters (and possibly up to trillions of parameters for the proprietary models). Input context window sizes varied from 8,192 tokens to one million tokens. Since the IG requires a longer context length than SPGs, we only used a subset of the models for the IG with a sufficient maximum context length window size. The open-source models were accessed via the cloud service inference provider Groq (Groq, 2025). We used a temperature of 0 for all LLMs to test classification capabilities.

3.2 Prompts Used

We added system-level prompts for the rules in the SPGs and the IG, and user-level prompts for the round-level information. The full prompts used can be found in Appendix B.

We used a text completion style prompt. Occasionally, models produced additional text; outputs were manually screened by the authors (to account

for reasoning) and reformatted to the preset prediction labels: `COOPERATE/DEFECT` and `INSPECT/NOT INSPECT`.

For the SPG prompt, to align with the human subjects study conducted by van Baar et al. (2022), we incorporated the LLM's guess and the ground truth (i.e., the player's actual choice) for past rounds into the context in the prompt. We conducted systematic experiments with a subset of LLMs to study the effect of different forms of feedback: (a) including only the ground truth, (b) including only the LLM's past guess, and (c) including both the ground truth and the LLM's past guesses (reported in Section 4). These experiments and analyses can be found in Appendix C.

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

In this research question, we contrasted LLM accuracy to human accuracy on SPGs, and investigated whether LLM accuracy on SPGs improves as the rounds progressed.

4.1 Difference Between Predicting Actions for Human and Artificial Motives

There was a significant gap in accuracy (two-tailed paired-samples t-test $t(149) = 22.0, p < 0.001$ (van Baar et al., 2022)) for SPGs between predicting actions for players with human motives (72%) vs. players with artificial motives (47%). As van Baar et al. (2022) also noted, people appear to predict players' actions by inferring their rational strategic motives (i.e., human motives). When players are modeled with counter-intuitive or irrational motives (i.e., artificial motives), humans struggle to reason about their intentions, and their accuracy tended to be close to chance. A purely statistical learning approach using a random forest classifier, however, showed no significant difference ($W = 307.5, p = 0.34$) in mean accuracy between artificial and human motives and predicted both with high accuracy. This was expected since the SPG motives are essentially step functions that can be modeled by a statistical machine learning model (Figure 2).

We found that all closed-source models had higher accuracies for SPGs with human motives than for SPGs with artificial motives, similar to 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 $t(33) = 15.77, p < 0.001$, Cohen's d $= 3.21$; GPT 4o: two-tailed paired-samples t-test $t(33) = 30.94, p < .001$, Cohen's d $= 7.18$; Claude Sonnet 4.5: Wilcoxon signed-rank test $W = 0.0, p < 0.001$; Gemini 2.5 Flash: Wilcoxon signed-rank test $W = 0.0, p < 0.001$), indicating that these LLMs were able to better predict the actions of players with human motives than those with artificial motives in an SPG. Note that the gap, compared to humans, is much larger for GPT 4o and Gemini 2.5 Flash, with the average accuracy for artificial motives being significantly lower than for human motives. This is interesting, as these LLMs have demonstrated similar accuracy to humans on SPGs with human motives, but much lower accuracy than humans on artificial motives, suggesting that the models may be overfitting to human motives. This could be the reason behind the much lower accuracy for SPGs with players with artificial motives. Although GPT 4o outperformed the newer GPT 4.1 on human motives, GPT 4.1 showed a smaller gap in accuracy between players with human and artificial motives, indicating that the newer GPT models may be slightly better at detecting patterns rather than overfitting on human motives from training data. GPT 4.1 also has a very similar gap in accuracy between human and artificial motives as in the human study, indicating higher alignment with humans.

The open-source Llama models we tested demonstrated a similar gap in accuracy between human and artificial motives as humans; Llama 3.3 (two-tailed paired-samples t-test $t(4) = 25.87, p < 0.001$, Cohen's d $= 5.58$), Llama 4 Maverick ($t(4) = 20.11, p < .001$, Cohen's d $= 5.44$) showed significant differences in accuracy between human and artificial motives. The newer of the two Llama models, Llama 4 Maverick, performed slightly better than Llama 3.3.

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 players with artificial motives. LLMs appear to mirror this behavior in SPGs. This preference for human motives impairs the ability of both humans and LLMs to infer artificial motives, implemented in this context as a step function. Sta-

Model Family	Model	n	Human		Artificial		Difference
			Mean (Std)	95% CI	Mean (Std)	95% CI	| Human-Artificial |
Statistical	Random Forest Classifier	34	85.57 (5.70)	[84.22, 86.93]	86.40 (4.28)	[85.38, 87.42]	0.83
Human (van Baar et al., 2022)	–	150	71.62 (10.54)	[69.93, 73.31]	46.55 (12.42)	[44.56, 48.54]	25.07
GPT	GPT 4.1	34	67.19 (7.47)	[65.42, 68.97]	**42.92 (7.44)**	[41.15, 44.69]	24.27
	GPT 4o	34	73.81 (5.97)	[71.80, 75.82]	33.0 (5.21)	[31.25, 34.75]	40.81
Claude	Claude Sonnet 4.5	34	71.97 (7.89)	[69.32, 74.62]	39.15 (5.93)	[37.16, 41.14]	32.82
Gemini	Gemini 2.5 Flash	34	**75.46 (5.31)**	[73.68, 77.25]	27.11 (4.47)	[25.61, 28.61]	**48.35**
Llama	Llama 3.3 70B	34	66.82 (7.39)	[64.34, 69.30]	30.97 (5.05)	[29.27, 32.67]	35.85
	Llama 4 Maverick	34	67.0 (6.91)	[64.68, 69.32]	31.53 (5.89)	[29.55, 33.51]	35.47

Table 1: Accuracy results (%) of open- and closed-source LLMs compared to human and statistical baselines on SPGs. Both humans and most LLMs perform higher on SPGs for players with human motives compared to players with artificial motives.

tistical models, such as random forest classifiers, on the other hand, perform better on SPGs for both players, and with both types of motives. This insight may be useful to future development of LLMs aiming to model human behavior.

4.2 Temporal SPG Accuracy

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

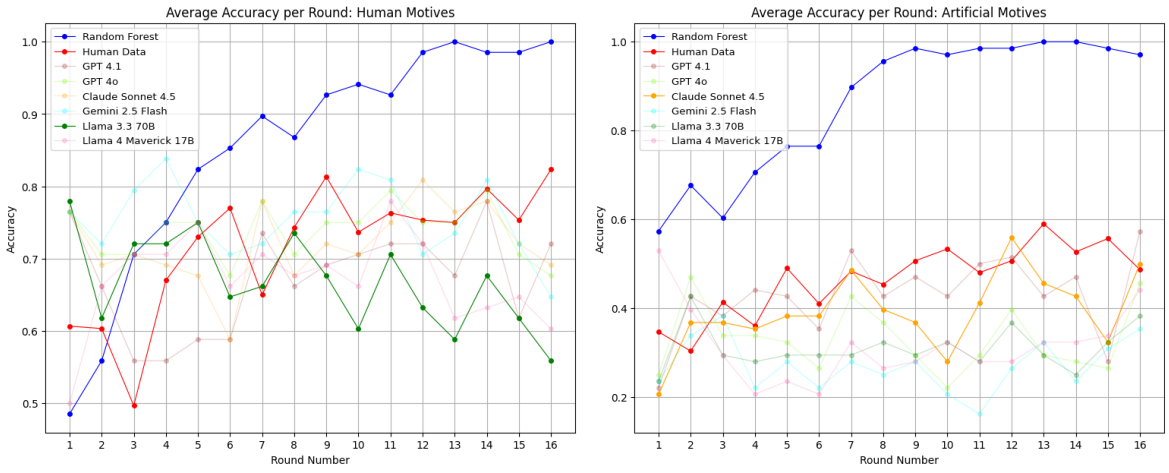
Figure 4 plots the temporal accuracy for LLMs from each model family with the highest average accuracy on SPGs for human motives. The statistical and human baselines follow a significant increasing trend as the rounds progress, demonstrating that average accuracy improves as more examples of past player actions are shown for an SPG for both human (human: $\tau = 0.59, Z = 3.15, p = 0.002$, random forest: $\tau = 0.89, Z = 4.80, p < 0.001$) and artificial motives (human: $\tau = 0.63, Z = 3.34, p = 0.001$, random forest: $\tau = 0.76, Z = 4.10, p < 0.001$). However, most LLMs do not follow an increasing trend in accuracy for SPGs for human (GPT 4.1: $\tau = 0.3, Z = 1.59, p = 0.11$, GPT 4o: $\tau = 0.03, Z = 0.09, p = 0.93$, Claude Sonnet 4.5: $\tau = 0.2, Z = 1.04, p = 0.30$, Gemini 2.5 Flash: $\tau = -0.15, Z = -0.77, p = 0.44$, Llama 4 Maverick: $\tau = -0.17, Z = -0.86, p = 0.39$) nor for artificial motives (GPT 4.1: $\tau = 0.33, Z = 1.74, p = 0.08$, GPT 4o: $\tau = -0.11, Z = -0.54, p = 0.59$, Gemini 2.5 Flash: $\tau = 0.04, Z = 0.18, p = 0.86$, Llama 3.3: $\tau = 0.20833, Z = 1.12, p = 0.26$, Llama 4 Maverick: $\tau = 0.22, Z = 1.14, p = 0.25$). Only two LLMs do show significant trends: Claude Sonnet 4.5 shows a significant increasing trend in accuracy for artificial motives ($\tau = 0.4, Z = 2.13, p = 0.03$). Conversely, Llama 3.3 shows a significant *decreasing* trend in accuracy for human motives ($\tau = -0.46, Z = -2.44, p = 0.01$).

Insight: The majority of LLMs we tested did not improve in accuracy in the Social Prediction Game with more observations of the Player's actions, indicating a lack of reasoning capabilities.

4.3 SPG Confidence

Finally, we analyzed model confidence with respect to SPG accuracy and the number of rounds observed in SPG (Table 2). van Baar et al. (2022) had participants predict the action they thought Player A would choose, and rate their confidence in that prediction on a scale of 0-100%. LLMs predict a probability distribution over the model token vocabulary, and with greedy decoding, the model output is the token(s) with the highest probability. This probability can be interpreted as the model's confidence in its output. Model confidence calibration is a highly explored area of research, with the focus on assessing models beyond accuracy and evaluating whether correct predictions have a higher confidence than incorrect predictions (Liu et al., 2025). We tested whether this was the case for human and LLM results by testing accuracy versus confidence. We saw for human motives, people's confidence increased with accuracy, but not for artificial motives. GPT 4o yielded similar results with humans for artificial motives, but none of the LLMs had similar results for human motives.

We then tested whether confidence increased with the number of rounds observed. More observations reveal the underlying pattern (Tenenbaum,



(a) (b)

Figure 4: (a) Accuracy (averaged over all iterations) of the SPG for each round for players with human motives. (b) Accuracy (averaged over all iterations) of the SPG for each round for players with artificial motives. For both motives, the human (red) and statistical model (blue) accuracy followed a statistically significant increasing trend as the rounds progressed.

1998) and so it is intuitive to expect higher confidence as the rounds progress. People's confidence increased with rounds for human motives, however only Gemini 2.5 Flash followed these results. For artificial motives, people's confidence decreased with rounds, and GPT 4o yielded similar results with humans. Gemini yielded opposite results.

Interestingly, only the human results for the human motives consistently aligned with the expected behavior for each experiment, suggesting that human motives are more intuitive to people. Based on the human results, GPT 4o and Gemini 2.5 Flash each had partial alignment with the human results.

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

We investigated LLM performance on the IG using only models with sufficient context window size. Risk-averse players always chose `WORK` in the IG, while greedy players always chose `SHIRK`, and the goal was to check if the model used what it observed about the opponent player's motive during the SPG to strategically inspect or not inspect the opponent (playing the Employee) in the IG. van Baar et al. (2022) found that humans demonstrated the capability to generalize the underlying motive to the IG (with statistical significance), which improved with higher accuracy on the SPG, indicating that people are able to generalize a discovered underlying motive to a new context. Table 3 shows that only GPT 4.1 and Llama 4 Maverick were able to correctly rank the risk-averse player as more likely to work compared to the greedy player, however the difference in rankings was not statistically significant (GPT 4.1: Wilcoxon signed-rank test $W = 214.0, p = 0.15$; Llama 4 Maverick: Wilcoxon signed-rank test $W = 267.0, p = 0.61$. On the other hand, Claude Sonnet 4.5 predicted the greedy player as being more likely to be working, which yielded a significant Wilcoxen signed-rank test ($W = 49.0, p = 0.001$). Although the rankings align somewhat with the underlying motive of the player, the amount of money was 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 were willing to pay to inspect risk-averse versus greedy players (GPT 4.1: Wilcoxon signed-rank test $W = 13.5, p = 0.57$; Claude Sonnet 4.5: Wilcoxon signed-rank test $W = 13.5, p = 0.28$; Llama 4 Maverick: Wilcoxon signed-rank test $W = 35.0, p = 0.77$). GPT 4.1 showed slightly higher willingness to pay for greedy players than risk-averse players; however, the difference was not statistically significant.

Insight: LLMs we tested were not able to generalize the discovered human motives to a new context of the Inspection Game, indicating that they did not discover the underlying motives in the Social Prediction Game.

Our results show that the tested LLMs were not

Data	Confidence Increases w/ Accuracy?		Confidence Increases w/ Rounds?	
	Human Motive	Artificial Motive	Human Motive	Artificial Motive
Human (van Baar et al., 2022)	Yes	No	Yes	No
GPT 4o	N/A	No	No	No
GPT 4.1	N/A	N/A	N/A	N/A
Gemini 2.5 Flash	N/A	N/A	Yes	Yes
	(a)		(b)	

Table 2: Model confidence analysis. (a) Results of Pearson correlation or Spearman's rank correlation coefficient correlation tests between confidence and accuracy. We report "Yes" if there is a significant ($p < 0.05$) positive relationship between the two variables, "No" if there is a significant negative relationship, and "N/A" if there is not enough evidence for any relationship. (b) Results of Mann-Kendall test of confidence over rounds in an SPG game. If there is a significant ($p < 0.05$) positive trend, we report "Yes", "No" if there is a significant negative trend, and "N/A" if there is not enough evidence for any trend. For full individual motive results, refer to Appendix A.2.

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 (N/A)	55 (N/A)	20 (9)	15 (9)
GPT 4.1	42.79 (14.2)	47.35 (17.25)	5.94 (9.26)	5.74 (9.33)
Claude Sonnet 4.5	56.12 (3.63)	51.59 (4.69)	1.18 (4.76)	3.5 (6.87)
Llama 4 Maverick	53.82 (14.66)	52.21 (13.35)	4.74 (9.3)	5.71 (10.12)

Table 3: Results of three recent LLMs 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 34 repetitions of IG.

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 SPGs. The increasingly long LLM context history may be contributing to the low generalization capabilities, with LLMs possibly forgetting past observed information.

6 Related Work

6.1 LLMs for Modeling Human Behavior

LLMs are increasingly being explored to model and simulate human behavior across various contexts, such as in sociological surveys or studies (Wang et al., 2025; Kang et al., 2023; Kolluri et al., 2025), social networks and multi-agent systems (Bougie and Watanabe, 2025; Park et al., 2023), and decision making (Tak et al., 2026; Wu et al., 2025; Jia et al., 2024). A growing body of research focuses on measuring alignment between LLM outputs and human behavior patterns, with particular attention on the risks of oversimplifying or misrepresenting the complexity of human decision-making (Abdurahman et al., 2024).

Recent work has established various methodologies for evaluating how well LLMs capture human behavior 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. Relative to MoCa, which examined static moral judgment tasks, our work focused on economic game action prediction where LLMs must infer the underlying motive from sequential game play to predict future opponent actions.

Aher et al. (2023) evaluated LLMs as proxies for human participants in behavior studies, finding some promising correlations but also systematic biases, especially in risk assessment and emotional decisions. Our work focused on LLMs' ability to model human strategic decision-making in validated games, predicting actions based on inferred motives rather than replacing human participants.

6.2 LLMs in Game Theory

Game theory provides a mathematical framework for modeling strategic interactions among ratio-

nal agents. It can be used to study human behavior in simplified, abstract scenarios known as economic games. These games have played a central role in behavioral economics and social science research for decades, revealing behavioral patterns in cooperation and competition (Sigmund, 2010; Poncela-Casasnovas et al., 2016; Ledyard et al., 1994). Recently, LLMs have opened new avenues for investigating decision-making behavior.

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 deviated from Rational Choice Theory in games like rock-paper-scissors, revealing issues with belief updating and payoff optimization. TMGBench further found that while LLMs performed 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 limited strategic reasoning.

These works primarily focused on LLM game-playing capabilities, however there has not yet been a systematic study of LLMs' ability to infer human actions. Exploring these fundamental skills for strategic decision-making is an important step towards understanding LLMs' capabilities and limitations in real-world human behavior modeling.

7 Conclusion

We explored two research questions regarding how LLMs compare to humans in determining underlying motives in SPGs and the IG. We gained insights into how LLMs appear to mimic human prediction of actions of players with human motives with higher accuracy compared to those of players with artificial motives. Furthermore, our results demonstrate that LLMs did not improve in accuracy as the rounds in the SPG progressed, unlike humans and a standard statistical model. Secondly, we found that LLMs were not able to generalize the learned motive in SPGs to a new context in the IG, unlike humans. This work identifies differences in the behavior of LLMs when modeling human behavior in the context of game theory. Our insights highlight the need for further research into human-AI alignment.

8 Limitations

Our SPG and the IG prompting methods do not employ mechanisms to verify whether models understand the game, unlike Fontana et al. (2025), who assessed rule comprehension and data parsing in an iterated Prisoner's Dilemma.

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

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.

9 Ethical Considerations

We are aware that running inference using multi-billion parameter LLMs can be costly to the environment. Our preliminary experiments show that smaller 3-13 billion parameter open-source models have a harder time responding in the correct format, and so future work can look into improving smaller models for these LLM evaluation studies.

Also, although LLMs may perform similarly to humans (e.g., GPT 4.1), they should not be considered replacements for human subject studies. As shown through our temporal analysis, LLMs may not be reasoning in a similar way as people.

Acknowledgments

Kaleen Shrestha is supported by the National Science Foundation CISE Graduate Fellowships under Grant No. 2313998. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7):pgae245.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 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.

Rudolf Avenhaus, Bernhard Von Stengel, and Shmuel Zamir. 2002. Inspection games. *Handbook of game theory with economic applications*, 3:1947–1987.

Nicolas Bougie and Narimawa Watanabe. 2025. Citysim: Modeling urban behaviors and city dynamics with large-scale llm-driven agent simulation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 215–229.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and 1 others. 2024. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. 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, pages 17960–17967.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. 2024. Nicer than humans: How do large language models behave in the prisoner's dilemma? *arXiv preprint arXiv:2406.13605*.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. 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, pages 522–535.

Groq. 2025. Groq cloud. Technical report, Groq.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E McNamara, and Deming Chen. 2024. Decision-making behavior evaluation framework for llms under uncertain context. *Advances in neural information processing systems*, 37:113360–113382.

Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. 2023. From values to opinions: Predicting human behaviors and stances using value-injected large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15539–15559, Singapore. Association for Computational Linguistics.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S Bernstein. 2025. Finetuning llms for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30084–30099.

John O Ledyard and 1 others. 1994. *Public goods: A survey of experimental research*. Division of the Humanities and Social Sciences, California Inst. of Technology.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117.

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.

Allen Nie, Yuhui Zhang, Atharva Amdekar, Chris Piech, Tatsunori Hashimoto, and Tobias Gerstenberg. 2023. Moca: Measuring human-language model alignment on causal and moral judgment tasks.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Julia Poncela-Casasnovas, Mario Gutiérrez-Roig, Carlos Gracia-Lázaro, Julian Vicens, Jesús Gómez-Gardeñes, Josep Perelló, Yamir Moreno, Jordi Duch, and Angel Sánchez. 2016. Humans display a reduced set of consistent behavioral phenotypes in dyadic games. *Science advances*, 2(8):e1600451.

Karl Sigmund. 2010. *The calculus of selfishness*. Princeton University Press.

Ala N Tak, Amin Banayeeanzade, Anahita Bolourani, Fatemeh Bahrani, Ashutosh Chaubey, Sai Praneeth Karimireddy, Norbert Schwarz, and Jonathan Gratch. 2026. Sparks of rationality: Do reasoning llms align with human judgment and choice? *arXiv preprint arXiv:2601.22329*.

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

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

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

Jeroen M van Baar, Matthew R Nassar, Wenning Deng, and Oriel FeldmanHall. 2022. Latent motives guide structure learning during adaptive social choice. *Nature Human Behaviour*, 6(3):404–414.

Haochuan Wang, Xiachong Feng, Lei Li, Yu Guo, Zhanyue Qin, Dianbo Sui, and Lingpeng Kong. 2024. Tmgbench: A systematic game benchmark for evaluating strategic reasoning abilities of llms. *arXiv preprint arXiv:2410.10479*.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. 2025. SocioBench: Modeling human behavior in sociological surveys with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26257–26289, Suzhou, China. Association for Computational Linguistics.

Weidong Wu, Qinlin Zhao, Hao Chen, Lexin Zhou, Defu Lian, and Hong Xie. 2025. Exploring the choice behavior of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5194–5214.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, and 1 others. 2024. Can large language model agents simulate human trust behavior? *Advances in neural information processing systems*, 37:15674–15729.

A SPG Full Results

A.1 SPG Accuracy Full Results

We present individual motive results for accuracy % value in Table 4. It is apparent for some LLMs and additionally for people, that greedy motive has a higher accuracy than risk-averse motive.

A.2 SPG Confidence Analysis Full Results

Tables 5 and 6 contain full results for the invidividual motives for the confidence-related experiments from Table 2.

B Prompts

Figures 5 and 6 show the prompts input to the LLMs 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.

C SPG Feedback Experiments

For a subset of LLMs, we run experiments with three different feedback conditions for past rounds in SPG: (a) including only ground truth, (b) including only LLM past guesses, and (c) including both ground truth and LLM past guesses. (c) is the feedback we use in the main work. Results are reported in Table 7.

For GPT 4.1, we see that including only ground truth (a) in the context history yields the highest accuracy for human motives. However, including both ground truth and LLM guesses gives the highest accuracy for artificial motive. Surprisingly,

Model Family	Model	n	Greedy	Risk-Averse	Inverse Greedy	Inverse Risk-Averse
			Mean (Std)			
Statistical	Random Forest Classifier	34	86.25 (4.68)	87.50 (0.00)	83.75 (5.00)	87.50 (3.95)
Human (van Baar et al., 2022)	–	150	87.00 (12.90)	56.25 (19.53)	45.08 (19.72)	48.00 (14.40)
GPT	GPT 4.1	34	68.2 (12.53)	66.18 (9.95)	43.57 (11.29)	42.28 (11.04)
	GPT 4o	34	81.62 (11.13)	65.99 (10.3)	29.6 (10.96)	36.4 (8.37)
Claude	Claude Sonnet 4.5	34	85.11 (10.06)	58.82 (12.87)	33.27 (8.12)	45.04 (9.07)
Gemini	Gemini 2.5 Flash	34	76.47 (6.44)	74.45 (8.76)	26.84 (7.03)	27.39 (6.25)
Llama	Llama 3.3 70B	34	77.94 (11.67)	55.7 (9.27)	19.3 (9.51)	42.65 (7.34)
	Llama 4 Maverick	34	72.24 (12.7)	61.76 (7.84)	24.45 (8.76)	38.6 (9.76)

Table 4: Full SPG accuracy results (%) for individual motives expanded from Table 1.

Data	Confidence Increases w/ Accuracy?			
	Greedy	Risk-Averse	Inverse Greedy	Inverse Risk-Averse
Human (van Baar et al., 2022)	YES r[16]=0.89, p<0.001	NO r[16]=-0.63, p=0.01	N/A r[16]=-0.84, p<0.001	N/A r[16]=-0.47, p=0.07
GPT 4o	N/A r[16]=0.01, p=0.96	NO r[16]=-0.54, p=0.03	NO r[16]=-0.50, p=0.047	N/A r[16]=-0.15, p=0.57
GPT 4.1	N/A r[16]=-0.04, p=0.87	NO r[16]=-0.63, p=0.01	N/A r[16]=-0.38, p=0.80	N/A r[16]=0.28, p=0.51
Gemini 2.5 Flash	N/A r[16]=-0.30, p=0.26	N/A r[16]=0.15, p=0.36	N/A r[16]=0.19, p=0.48	N/A r[16]=0.06, p=0.82

Table 5: Full confidence versus accuracy correlation results for individual motives, expanded from Table 2.

Data	Confidence Increases w/ Rounds?			
	Greedy	Risk-Averse	Inverse Greedy	Inverse Risk-Averse
Human (van Baar et al., 2022)	YES = 0.89, Z = 4.78, p <0.001	NO = -0.88, Z = -4.69, p <0.001	NO = -0.81, Z = -4.34, p <0.001	NO = -0.51, Z = -2.70, p = 0.01
GPT 4o	N/A = -0.37, Z = -1.94, p = 0.05	NO = -0.65, Z = -3.47, p = 0.001	N/A = -0.37, Z = -1.94, p = 0.05	NO = -0.38, Z = -2.03, p = 0.04
GPT 4.1	N/A = 0.2, Z = 1.04, p = 0.30	N/A = -0.25, Z = -1.31, p = 0.19	N/A = -0.18, Z = -0.95, p = 0.34	N/A = -0.25, Z = -1.31, p = 0.19
Gemini 2.5 Flash	YES = 0.43, Z = 2.30, p = 0.02	YES = 0.42, Z = 2.21, p = 0.03	YES = 0.42, Z = 2.21, p = 0.03	YES = 0.4, Z = 2.12, p = 0.03

Table 6: Full confidence trend results for individual motives, expanded from Table 2.

Model	Feedback Mode					
	a		b		c	
	Human	Artificial	Human	Artificial	Human	Artificial
GPT 4.1	**78.22 (4.88)**	36.76 (5.87)	71.6 (6.71)	27.39 (6.57)	67.19 (7.47)	**42.92 (7.44)**
LLama 3.3 70B	**68.75 (6.11)**	31.43 (5.67)	62.41 (10.94)	**38.05 (10.6)**	66.82 (7.39)	30.97 (5.05)

Table 7: Accuracies (%) in SPG for each type of feedback (a) including only ground truth, (b) including only LLM past guesses, and (c) including both ground truth and LLM past guesses.

not having past knowledge of ground truth gives a higher accuracy for human motives than including both. This is not the case for artificial motives, where having both the LLM guess and the ground truth has the highest accuracy. For Llama 3.3, we see that like in GPT 4.1, SPG accuracy is highest for condition (a) for human motives, but we see artificial motives SPG accuracy is highest for (b). It is worth studying in further detail why human motive SPG performance is highest when only ground truth signal is given.

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. Saving money is your goal. 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.