Rethinking LLM Social Intelligence: Evidence from Multi-Agent Bartering Scenarios

Anonymous submission

Abstract

Large-Language models (LLMs) have demonstrated wide ranging capabilities on a multitude of variegated tasks. We use a novel bartering scenario where individuals must propose and convince prospective trade partners in order to make profitable trades. Successful bartering requires the development of a social strategy for identifying and convincing potential trade partners. We compare the performance of groups of human traders to the performance of groups of LLM-based agents in different versions of the base trade scenario that motivate cooperation, competition, or independence. Our results indicate that LLM agents, unlike humans, neither develop condition-specific strategies nor do their strategies evolve over the course of a trade session. Whereas humans quickly and universally develop simple but effective pro- and anti-social strategies, LLM agents demonstrate limited in-context learning in this scenario. We conclude that prior research using traditional economic games to demonstrate social strategy development by LLMs may actually reflect pre-training on these games.

Introduction

Large-language models (LLMs) hold great promise as a tool for simulating human behavior and decision-making in a wide variety of fields including psychology, politics, sociology, and economics (Manning, Zhu, and Horton 2024; Ziems et al. 2024; Ke et al. 2025). LLM-controlled agents have been proposed as a means for developing research, educational, and medical assistants (Schmidgall et al. 2025; Kweon et al. 2025; Vrdoljak et al. 2025) because of their ability to interact with people in a natural and human-like manner. Recent research has begun to investigate whether LLM agents behave like humans in simulated and, in some cases, complex environments (Lu et al. 2024b; Gürcan 2024). While many studies test whether agents can achieve goals based on a given task, a key question is whether LLM-controlled agents can replicate not just the outcomes of human decisions, but the nuanced, socially-aware reasoning that underlies them.

In this work, we investigate the development and use of social strategies by LLM-controlled agents acting in a multi-agent bartering setting. Bartering is an attractive environment for investigating social strategies because the primary motivated task involves identifying and convincing potential trade partners and/or deciding whether an offered trade is advantageous. Moreover, a barter can be cooperative when trades are mutually beneficial to both traders, or competitive when one must compete for buyers (Von Neumann and Morgenstern 1947).

We therefore designed a controlled environment where groups of participants seek to negotiate and trade items over a 30-minute session in order to maximize the value of items they possess. We alter the global incentive structure across different experimental conditions, promoting cooperation (collective success); competition; or pursuit of independent goals. This allows us to evaluate not only baseline strategic competence but also the ability to adapt behavior to varying social contexts.

Our experimental design moves beyond short-horizon testing prevalent in the literature (Xie et al. 2024; Sreedhar and Chilton 2024) and evaluates agent performance in a dynamic setting that requires long-term planning, strategy, communication, negotiation, and social awareness. Moreover, traditional behavioral economic paradigms—such as the Prisoner's Dilemma, Ultimatum Game (Harsanyi 1961), and Trust Game (Berg, Dickhaut, and McCabe 1995)—have extensive analyses of strategy widely available online, much of which are likely part of LLM training data. An agent's "strategic" behavior in these games may not reflect genuine in-context reasoning but rather the retrieval of learned patterns which can lead to inflated performance assessments and a misleading perception of their true capabilities(Xu et al. 2024). Unlike well-documented economic games, the social strategies humans use when bartering are not widely available online, and general information about bartering is unlikely to aid an agent in our paradigm.

Our work addresses the following research questions:

- **RQ1** How does the performance of groups of LLM-controlled agents on a bartering task compare to that of human groups?
- **RQ2** Do LLM-controlled agents and human traders use similar social strategies?
- **RQ3** Do LLM-controlled agents adapt their trading strategies in response to context, such as the type of incentive or time pressure?

Our primary contribution is a direct and controlled comparison of the social strategies and interactive behavior between groups of human participants and groups of LLM

agents within the simulated task environment. Our findings suggest that while LLMs are instrumentally rational in bartering environment, they fail to adapt to social context–in our case, different incentive structures–a key differentiator from human intelligence. This finding is a departure from related work testing LLMs in different game theoretic settings, e.g.,(Mao et al. 2023; Xie et al. 2024). The remainder of this paper first summarizes the related work before detailing the methods used and presenting the results from controlled bartering experiments. Anonymized code and data from this work are shared[1].

Related Work

Our work is situated in the intersection of multi-agent systems, behavioral economics, and LLM evaluation. We first review the use of LLMs to simulate human behavior and examine the common approaches used to evaluate strategic reasoning.

LLMs for simulating human behavior

Several studies have explored the potential of LLMs to simulate human behavior in experimental settings (Bara, CH-Wang, and Chai 2021; Nebel, Schneider, and Rey 2016; Szymkiewicz 2022; Park et al. 2023; AL et al. 2024). LLMs were shown to be capable of simulating experimental subjects in economic research by providing them with endowments, information, and preferences Horton (2023). Similarly, studies demonstrated that when prompted with demographic information, LLMs could act as effective proxies for specific human sub-populations Argyle et al. (2023); Aher, Arriaga, and Kalai (2023). Such findings were extended to simulate diverse populations Aher, Arriaga, and Kalai (2023), while more recent research showed that LLMs can be used to synthesize sub-rational human decision making process, that deviates from perfect rationality Coletta et al. (2024). These society-level simulations hint that LLM-controlled agents may be capable of developing complex, context-dependent social strategies

LLMs in game theoretic environments

A common approach to evaluating the strategic capabilities of LLM agents involves testing them within controlled game environments, often drawing from game theory. LLM agents have been evaluated in complex world simulations like Minecraft Hu et al. (2024), text-based strategy games (Huang et al. 2024), and canonical economic games such as the Ultimatum Game (Sreedhar and Chilton 2024). In these settings, models like GPT-4 have demonstrated sophisticated capabilities, including negotiation Abdelnabi et al. (2023), engage in sophisticated social deduction, understanding and predicting deceitful behavior effectively O'Gara (2023), and even long-term planning that can exceed human performance Mao et al. (2023); Duan et al. (2024).

LLM-based multi-agent systems

Beyond single-agent evaluations, multi-agent systems based on LLMs are being increasingly used to automate human workflows for complex tasks such as coordinating LLM software engineering teams Hong et al. (2023) to serving as llm scientist Lu et al. (2024a), and in legal systems Li et al. (2024). However, many of these studies focus on defining specific predefined agentic roles and focus on task completion and operational efficiency, often failing to address the impact of automating inherently human social processes. Our work provides a contrasting analysis by highlighting what is lost in terms of emergent social roles and adaptive group dynamics when replacing humans with LLM-based agents. This aligns with a growing perspective that LLMs should be viewed as powerful tools to augment human capabilities, rather than as direct replacements for them (Van Rooij et al. 2024).

Synthesizing the current literature, while LLM agents demonstrate capabilities in planning, simulating certain human behaviors, and performing in various game settings, significant gaps remain in understanding their proficiency in complex, dynamic multi-agent interactions. Much prior work evaluates agents in relatively short-horizon tasks or canonical games (e.g., Trust Game, Ultimatum Game) which may be present in the models' training data, potentially inflating perceived performance or alignment. Furthermore, evaluations often utilize environments with predefined structures or communication protocols, limiting the assessment of agents' ability to autonomously develop emergent strategies, coordinate effectively, and adapt socially in less constrained, longer-duration scenarios. Our work aims to address these gaps by introducing a novel, dynamic task environment designed to understand these specific capabilities through direct comparison with humans.

Methods

Our experiments used a series of bartering scenarios embedded within a customized Minecraft environment. Traders begin with a set of items in their possession as well as a list of *desired items* they will seek to obtain over the course of the experiment through iterative trade with willing traders. We conducted two classes of experiments with: (1) human traders and (2) LLM-controlled agents. The environment, rules and tasks were identical for humans and agents.

Experimental environment All experiments were conducted using a modified Minecraft server built with Spigot (Szymkiewicz 2022), an open source version of Minecraft that provides an API for modifications to the Minecraft world. Minecraft (Mojang Synergies AB 2022; Bara, CH-Wang, and Chai 2021; Engelbrecht and Schiele 2013, 2014) was chosen because it supports player interaction, communication, and item exchange in an observable, controllable, and recordable fashion. Minecraft is also intuitive and familiar to our participants, who were required to have experience playing Minecraft and a local copy on their computer to participate. We developed the plugin which allowed humans and LLM agents to trade and communicate in real-time during experiments.

At the start of the experiment session, a character was spawned for each participant in an enclosed Minecraft village. Each participant was randomly assigned a profession

[1]https://anonymous.4open.science/r/23B1

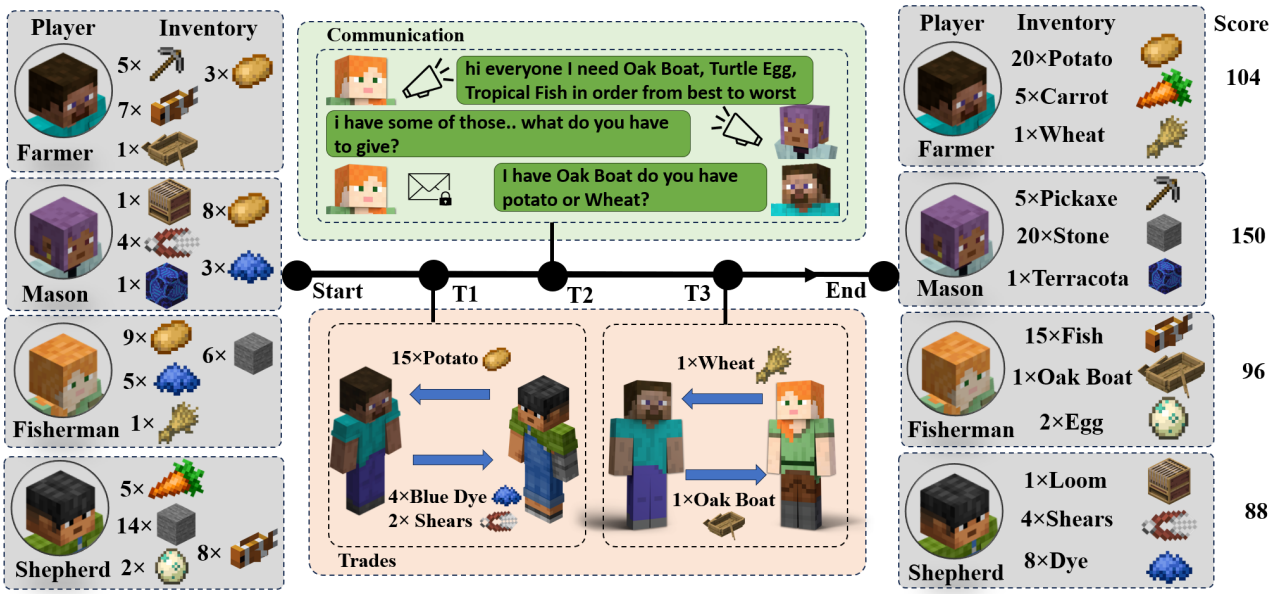

Figure 1: An overview of the experimental bartering scenario. Participants begin with randomly assigned professions and inventories (left), creating initial resource disparities. Throughout the session, they use global and private chats as communication modes and a trading system to exchange items (center). The right panel shows a sample outcome, where items have been successfully reallocated to align with profession-specific goals, resulting in increased scores.

from the default villager professions in Minecraft (e.g., fisherman, blacksmith) and received an initial random inventory of items to trade.

Each profession had a predefined set of desired items with a point value for each denoting its desirability. Possessing these items increased the participant's score. For instance, a fisherman's score increases if they add an oak boat to their inventory, but obtaining a boat would not increase the score of a blacksmith. Item values were structured into three tiers based on their rarity: Tier 1 included 3 common item types (20 occurrences each, 1 point each); Tier 2 included 2 somewhat more rare item types (10 occurrences each, 3 points each); and Tier 3 included 1 very rare item type (3 occurrences, 10 points each) (see the Supplemental Materials for a complete list of professions and items). A participant's score at any time was the sum of the values of the profession-relevant items currently held. The maximum possible score was 150. Crafting items and other means of acquiring items were disabled. Hence, participants could only increase their score by trading with others. Participants could communicate using either a global chat visible to all participants, or a private messages visible only to the sender and recipient.

Throughout the experiment, participants could easily view their own current inventory and score. However, participants could not directly see the inventories, scores, professions, or desires of other participants.

Each experiment session lasted 30-minutes. Upon ending the final scores were calculated based on the value of profession's relevant items held by each participant. Each trader's score, their total gain over the session, and their rank relative to other participants was displayed in the chat. Session logs automatically captured all trade information and all chats.

Trading Trades were initiated by one participant sending a *Trade Request* to another participant specifying the items being offered and items requested for trade. The recipient of a trade request could accept or decline the request. The trade remained pending until the recipient accepted or declined the trade or the initiator canceled the request. A participant could have multiple pending requests with different participants, but only one pending request per pair of participants. A trade request automatically failed if either participant lacked the necessary items required to complete the trade. Successful trades resulted in the exchange of offered and requested items between inventories.

Incentive structures To systematically vary the incentives motivating participants in each condition, $5 bonuses were offered as follows. *Competitive Condition:* Only the highest-scoring participant at the end of the session received the bonus. *Cooperative Condition:* All participants received the bonus only if each of them achieved a score greater than or equal to 120. *Baseline/Independent Condition:* Any participant who obtained a score greater than or equal to 100 points received the bonus.

Human subjects experiments

Human participants were recruited from the university campus using flyers and online posting on the university's subreddit and student discord servers. Participants were required to be at least 18 years of age, own Minecraft Java Edition, have prior experience playing Minecraft. They were compensated $10 for participating in a trade session of 30 minutes with the opportunity to win an additional $5 bonus. Participants were scheduled in groups of 5-7 for online ses-

sions. This study was IRB approved.

A total of 109 human subjects (73% male; 23.4% female; 3.6% did not disclose) participated. Seven cooperative, 7 competitive and 6 baseline sessions were conducted. Each session began on Zoom with an explanation of the study, consent forms, and instructions detailing the task, objectives, interface, and the specific bonus condition for their session, before joining the Minecraft server.

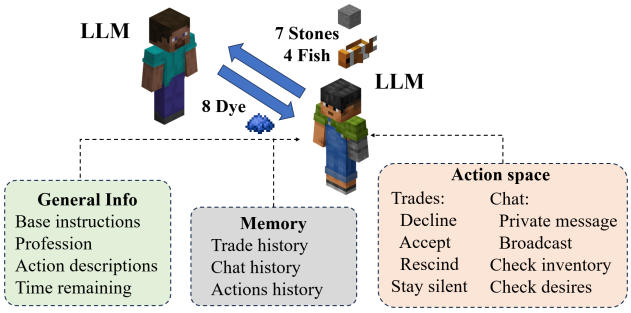

Figure 2: The architecture of LLM agent showing components of its information inputs, and available actions.

LLM agents experiments

A total of 120 agents participated. Two versions of Open AI's GPT-4o (2024-05-13 and 2024-08-06) (OpenAI 2025) and Google's Gemini-1.5-pro-002 (Team et al. 2023) were used. Eight sessions were conducted for each condition. Each agent received its initial information about its profession, inventory, and desired items, analogous to humans. Public information (time remaining, global chat, reminders) and private information (inventory, private messages, desires) was accessible either as agentic memory or via tool calls. Agentic memory stored information in chronological order, with new entries about the agent's activities appended as the session progressed. As a result, a complete record of all trades, conversations, and tool calls were retained. We used consistent hyperparameters for LLM generation with temperature 0.7, and TopP of 0.8 across all agent sessions. The LLM instance was refreshed across each session to prevent data leakage across sessions. All experimental procedures (duration, random assignment of professions and inventory, scoring, reminders, and memory) were identical to human sessions. Agents operated within the models' context window, each experiencing a progressively increasing buffer time ranging from 0 to 300 seconds to process information. This buffer duration was sufficient for the 30-minute sessions. Once any agent reached the 300-second limit, the buffer was reset to 0 for all agents.

LLM agents used the same mechanisms to communicate and trade as humans. Humans and agents could propose a trade, accept a trade, decline a trade, cancel a trade, list pending trades, check their score, desired items and inventory, and communicate using the global chat or private messages.

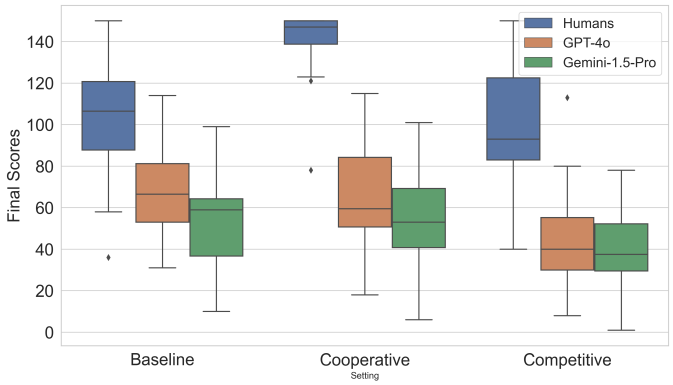

Figure 3: Comparison of final scores across participant groups and incentive condition.

Results

We analyzed logged data from both humans' and LLM agents' sessions. Key metrics include final scores, trade counts (initiated, accepted, declined, failed), and communication channel usage (global vs. private messages). We compared these metrics across participant groups (humans vs. GPT-4o and Gemini) and across incentive conditions (competitive vs. cooperative vs. baseline). We first establish that the LLM agents exhibit a baseline of goal-oriented behavior, and then we demonstrate how their simple form of rationality fails to align with the complex adaptive social strategies used by human participants.

Instrumental rationality of LLM agents

We first establish baseline behavior of LLM agents. We find that they consistently act as functional, goal-oriented participants demonstrating instrumental rationality—i.e., understanding of stated objectives and the ability to take logical actions in pursuit of the same (Weber 1978). Across all conditions, agents consistently used the available tools to pursue their stated objectives. For instance, agents frequently used the *check_desires* and *query_inventory* actions an average of about 7 and 14 times per session, respectively, demonstrating consistent information-seeking behavior. Furthermore, agents are proactive by initiating a total of 2855 trades across 72 sessions using the *propose_trade* action were for item exchange. They also demonstrated rational evaluation of incoming offers, by gaining 3.2 points per accepted trade. Taken together, this evidence demonstrates that agents successfully grasped the task's core mechanics and consistently acted to increase their scores.

Humans outperform agents across all experimental conditions

Across all conditions, human participants achieved significantly higher final scores than LLM agents. The overall mean scores for humans ($M_h = 115.4, SD = 31.1$) was nearly double that of LLM agent groups ($M_{LLM} = 55.7, SD = 22.1$), a statistically significant difference ($t(142) = 18.66, p = 4.8 \times 10^{-40}$) (see figure 3).

Table 1: Mean (SD) Total trades initiated by participant group and incentive condition.

Trader	Competitive	Baseline	Cooperative	Overall	Total
Humans	15.54 (9.94)	14.63 (6.89)	14.62 (6.81)	14.38 (7.95)	1568
gpt-4o-2024-05-13	9.2 (5.1)	12.5 (9.27)	11.73 (6.45)	11.13 (7.24)	1314
gpt-4o-2024-08-06	5.74 (4.61)	7.65 (5.03)	6.62 (4.91)	6.67 (4.87)	774
gemini-1.5-pro-002	4.57 (3.34)	5.25 (3.86)	6.28 (4.4)	5.36 (3.91)	557

The performance gap was most pronounced in the cooperative condition, where human groups effectively coordinated to reach near-optimal scores ($M_h = 141.6, SD = 13.9$), with 85.7% of human groups successfully coordinated securing the collective bonus, with the only exception involving a session with a non-responsive participant. In contrast, no LLM agent group met the collective goal. LLM agents performed comparatively poorly ($M_{LLM} = 58.7, SD = 24.9$), $t(109) = 22.88, p = 2.0 \times 10^{-43}$).

Condition-specific strategy adaptation

Human participants After an initial learning period, human participants developed condition-specific trade strategies. In the cooperative condition, they traded significantly higher-valued items early in the session ($M_{valtrd} = 28.8, SD = 32.8$ in the first ten minutes) as compared to later in the session ($M_{valtrd} = 10.9, SD = 18.1$ in the final ten minutes), $t(123) = 4.41, p = 2.0 \times 10^{-5}$. They tended to use broadcast messages to advertise their desire to trade to all members of the group ($M_{brd} = 89.6\%, SD = 13.1$).

In the competitive condition, humans traded similar-valued items early in the session ($M_{valtrd} = 11.7, SD = 19.1$ in the first ten minutes) as compared to later in the session ($M_{valtrd} = 14.3, SD = 21.2$ in the final ten minutes), $t(127) = 2.11, p = 0.036$. They tended to use private messages to seek out trade partners and fewer broadcast messages ($M_{brd} = 76.5\%, SD = 32.4; M_{pvt} = 23.5\%$).

Notably, human trading strategies tended to evolve as the session went on. In the baseline condition, for example, an average of 0.59 ($SD = 0.7$) trades were made in the first 10 minutes of the session whereas an average of 1.17 ($SD = 1.3$) trades were made in the final third of the trade session. In this condition, participants learned that they needed to increase trading in order to win the bonus.

LLM agents In contrast to humans, LLM agents did not adapt their strategies based on the trading condition or over the course of the session. LLM agents consistently traded higher-value items early in the session, regardless of condition ($M_{valtrd} = 9.43, SD = 11.4$ in the first 10 minutes vs. ($M_{valtrd} = 2.18, SD = 4.78$ in the final ten minutes), $t(212) = 7.32, p = 5.0 \times 10^{-12}$. Similar behavior was observed in the cooperative condition, ($M_{valtrd,early} = 11.06, SD = 13.52$ vs. $M_{valtrd,late} = 6.29, SD = 11.01, t(324) = 3.55, p = 4.3 \times 10^{-4}$) and baseline condition ($M_{valtrd,early} = 14.41, SD = 13.53$ vs. $M_{valtrd,late} = 4.48, SD = 8.23, t(277) = 8.08, p = 1.9 \times 10^{-14}$). LLM agents also used the same channel to communicate with potential trade partners regardless of the condition ($M_{brdcst,comp} = 47.46\%, SD = 42.0; M_{brdcst,coop} = 47.04\%, SD = 42.2; M_{brdcst,base} = 55.01\%, SD = 42.88$).

Trading activity

Human participants engaged in significantly more trading activity compared to LLM agents. Table 1, which summarizes the mean and standard deviation values of number of trades across the different participant groups and trading conditions, shows humans initiated an average of 14.9 trades ($SD = 8.0$) across all conditions whereas LLM agents initiated an average of 7.8 trades ($SD = 6.1$), $t(144) = 8.42, p = 3.4 \times 10^{-14}$.

LLM agents, however, *accept* a significantly higher percentage of initiated trades–that is to say, a significantly higher percentage of initiated trades are successful amongst LLM agents' ($M_h = 44.7\%, SD = 21.6$, vs. $M_{LLM} = 56.0\%, SD = 28.1$), $t(222) = -4.3, p = 2.5 \times 10^{-5}$.

Hoarding

Hoarding or *cornering the market* is a trade strategy that manipulates the market by denying competitors access to valuable resources (Putniņš 2012). In our experiment, non-valuable items held by a participant directly reduce the potential score of other participants who need that item. We measure hoarding, per player, as the total point value of items held by a player at the end of the game which have no value to them.

Human participants used hoarding strategically, with context-specificity. Human players hoarded significantly more in the competitive condition per trade ($M = 9.04, SD = 7.85$) than in the cooperative condition ($M = 5.18, SD = 3.39$), $t(37) = 2.2, p = 0.034$. This suggests hoarding may have been advantageous in competitive settings, where blocking others from obtaining desired items could prevent them from winning. In cooperative settings, reduced hoarding likely enabled all participants to accumulate enough points to reach the threshold and earn the bonus.

In contrast, LLM agents in the cooperative condition hoarded significantly more per trade ($M = 5.74, SD = 3.57$) than those in the competitive condition ($M = 4.14, SD = 4.06$), $t(128) = -2.38, p = 0.019$. This counterproductive behavior undermined cooperation, as agents failed to help one another meet the threshold score.

Trade arbitrage

Trade arbitrage or *multilateral trades* are strategic sequences of trades where a participant may acquire an item of no

Table 2: Mean (SD) of frequency of engaging in trade arbitrage across groups in different trading conditions.

Trader	Competitive	Baseline	Cooperative	Overall
Humans	4.73 (5.22)	3.69 (3.54)	3.03 (2.31)	3.78 (3.89)
gpt-4o-2024-05-13	2.0 (1.07)	3.13 (2.7)	2.17 (1.5)	2.53 (2.05)
gpt-4o-2024-08-06	1.0 (0.0)	1.62 (0.96)	1.33 (0.65)	1.45 (0.81)
gemini-1.5-pro-002	1.0 (0.0)	1.25 (0.5)	1.67 (1.03)	1.42 (0.79)

value with the intention of later trading the item to someone seeking the item. Use of this strategy demonstrates forward strategic planning. Formally, an intermediary trade by participant p involving item i occurs if:

1. Participant p acquires quantity $x > 0$ of item i via a trade Tr_1 at time t_1, where $V_{ik_p} = 0$.
2. Subsequently, at time $t_2 > t_1$, participant p successfully trades away at least a portion of the acquired quantity x of item i in a trade Tr_2 receiving in return a set of items R_2 that contains at least one item j which is valuable to them ($\exists j \in R_2$ such that $V_{jk_p} > 0$).

Given a trader from a particular participant group in a particular trading scenario has been observed to perform trade arbitrage, their mean and standard deviation of the count of such instances throughout the trading session is summarized in Table 2. **Human participants frequently engaged in trade arbitrage.** Specifically, 69 out of 109 (63.3%) participants engaged in trade arbitrage at least once, leveraging the strategy in a total of 262 trades. Human traders enjoyed relatively substantial score gain through arbitrage ($M = 12.47, SD = 11.77$).

While, 98 out of 360 (27.22%) LLM agents employed trade arbitrage in a total of 201 trades. Agents achieved significantly lesser score gains through arbitrage than their human counterparts ($M = 4.99, SD = 5.28$), $t(382) = 9.17, p = 2.89 \times 10^{-18}$.

Trade fairness

Trade fairness is calculated for each accepted trade, $Tr_{p \to p'}$, initiated by p and accepted by p'. Let $O_{p \to p'}$ be the set of *(item, quantity)* pairs offered by p to p', and $R_{p' \to p}$ be the set of *(item, quantity)* pairs offered by p' to p (which p requests). Assuming a global, profession-independent point value GV_j for each item type j, the trade fairness metric is defined as the ratio of the total objective point value of the items requested to the total objective point value of the items offered.

$$F_{Tr} = \frac{\sum_{(j,qty_j) \in R_{p' \to p}} qty_j \cdot GV_j}{\sum_{(j,qty_j) \in O_{p \to p'}} qty_j \cdot GV_j} \quad (1)$$

A fairness ratio greater than 1 indicates that the initiator gained more points from the trade, whereas a ratio less than 1 suggested that the recipient gained more points. Ratios significantly different from 1 may indicate strategic negotiation tactics (e.g., extracting surplus value if greater than 1, offering concessions if less than 1) or reflect underlying power dynamics between traders.

Overall, LLM traders ($Mean_{LLM} = 1.38, SD = 1.89$) **were more fair than humans** ($Mean_h = 2.11, SD = 4.08$) **across all conditions when participating in a trade exchange**, $t(776) = 4.28, p = 2.07E-05$. This effect is most pronounced in the cooperative condition where trade initiators in human studies ($Mean_h = 2.52, SD = 4.9$) profited much more than LLM-based trade initiators ($Mean_{LLM} = 1.34, SD = 1.85$), $t(271) = 3.55, p = 4.49E-04$ through each accepted trade. While in the competitive condition, humans ($Mean_h = 1.98, SD = 3.99$) and LLM agents ($Mean_{LLM} = 1.6, SD = 2.27$) are similarly fair. This result suggests that initiating a trade was more advantageous for human traders compared to LLM-based agents. In the cooperative condition, human trade initiators received, on average, 2.5 times more than what they offered, indicating that their efforts to communicate needs were recognized and met with generous responses from others. In contrast, LLM agents consistently exhibited more balanced, fair exchanges across all conditions, showing little variation in behavior based on social or strategic context.

Discussion

Our results reveal a gap between human social intelligence and the capabilities of the LLM agents. While LLM agents demonstrated instrumental rationality, they exhibited strategic inflexibility. Human participants dynamically adapted their strategies which aligns with observations from behavioral economics that human decision-making cannot be reduced to simple, utility-maximizing logic especially under uncertainty (Simon 1955; Tversky et al. 1982; Simon 1997). In contrast LLMs used a rigid one-size-fits-all approach that was often counter-productive. While LLM agents behaved like a textbook *homo economicus* assumed in classical game theory (Fehr and Gächter 2000; Von Neumann and Morgenstern 1947), human participants demonstrated a more sophisticated, socially-aware rationality. We argue this failure stems from a lack of deep social cognition, manifesting as an inability to model others' intentions, a reliance on surface-level heuristics, and a failure to engage in long-term strategic planning.

Absence of adaptive strategy

The central finding of our work is strategic inflexibility of LLM agents. Human participants clearly demonstrated a functional theory of mind by modeling the intentions of others based on the shared context. In competitive condition, human behavior by hoarding resources and communicating privately can be seen as a competitive sabotage, where players incur a small cost to impose a larger cost on rival. Conversely, in the cooperative condition, humans shifted to open communication and engaged in unfair trades that benefited

the group, understanding the importance of collective success.

The LLM agents, by contrast, failed to model these dynamics. Their behavior suggests they operate on simple, surface-level heuristics rather than a deep understanding of these social concepts. This was further evidenced by their temporal insensitivity; unlike humans, who adjusted their trading pace based on time pressure, the agents followed a static pattern of behavior throughout the session.

Strategic myopia and surface-level heuristics

The specific behaviors we analyzed further point to the agents' cognitive limitations. Their rigid adherence to fair trades, where value is exchanged equally, is indicative of a learned, generic rule for negotiation that ignores the social utility of a trade. However, humans understood that fairness is context-dependent and is deeply intertwined with intentions, reciprocity, and social context (Fehr and Schmidt 1999; Fehr and Gächter 2000). This contrast highlights the agent's limitation in understanding is limited and their strategies constrained by simple rules that prevent them from grasping higher-order concepts like trust and collective benefit.

Similarly, the infrequent use of trade arbitrage among agents points to a reactive decision-making process. They failed use a forward-looking reasoning required to see an item not as an end in itself, but as a means to a future, more valuable trade a clear sign of limited strategic depth which aligns with recent findings that show that LLMs over-exploit known, high-reward actions at the expense of long-term exploration and planning (Schmied et al. 2025). This suggests that even with a chronologically-ordered memory of events, the models are unable to synthesize this information into a coherent, long-term strategy. The counter-productive hoarding in the cooperative condition is the another example of this, where an action that might be rational in a competitive context directly undermined the collective goal.

Limitations and Future Work

Our study has several limitations that provide clear directions for future work. First, a key limitation stems from inconsistency in experiment design. One of the models tested, an earlier snapshot of GPT-4o (*gpt-4o-2024-05-13*) was implemented with a constraint that only allowed single item trades. This design choice likely reduced its performance and makes it challenging for a direct comparison to LLM agents and humans.

Second, the human participants in our experiments are from a university student population which may not be representative of the general population. Third, the LLM agents used in this work were implemented in a simple zero shot design. While this provides a baseline, more sophisticated agent architectures with explicit planning, reflection modules may yield better LLM agent performance. Finally, the bartering environment is a simplification of complex social interactions in real world. This single session design with no long term history between the sessions may limit scope of emergent social phenomenon.

In contrast, humans display hoarding behavior exclusively in competitive scenarios, where external pressures or constraints may influence their trading strategies. This suggests that while agents are predisposed to hoarding as a general behavior, humans are selective and only resort to hoarding under specific conditions that likely involve heightened competition or scarcity of desired resources. This distinction highlights a fundamental difference in decision-making and adaptability between humans and agents in trading environments.

A potential future direction for this work involves extending it into a hybrid framework where multiple agents, each powered by different LLMs, collaborate to complete bartering tasks. This approach could help investigate whether such agents demonstrate increased collaboration or trust across models, or whether they tend to align more closely with agents of the same kind. Another promising direction would be to integrate both human participants and LLM-based agents within the same environment to examine how trust and cooperation emerge in mixed human-agent bartering interactions.

Conclusion

This study explored the performance of LLMs in a complex bartering scenario that required strategic social interaction. A custom virtual environment was built in Minecraft, where human participants engaged in item-based trades. The same experimental setup was later used with three LLM-based agents: *gpt-4o-2024-05-13*, *gpt-4o-2024-08-06*, and *gemini-1.5-Pro-002*. By comparing the behaviors of human participants and LLM agents across cooperative, competitive, and independent trading conditions, we observed key differences in adaptability and strategic development. Human traders quickly adopted condition-specific strategies and adjusted their behavior as the session progressed. They also demonstrated a range of complex behaviors that changed depending on the trading condition—for example, hoarding in the competitive setting, fair trade practices in the cooperative setting, and facilitating intermediary trades to support broader exchanges. In contrast, LLM agents showed limited behavioral variation and minimal learning throughout the sessions. These differences were reflected in the final outcomes, with humans consistently outperforming LLM agents in both final scores and net worth gains over the 30-minute trading period. Overall, our findings suggest that current LLMs face significant limitations in developing and adapting social strategies in dynamic, unfamiliar environments. Furthermore, prior demonstrations of strategic behavior in LLMs may be influenced by pretraining on structured economic games, rather than true in-context learning. These results highlight ongoing challenges in designing autonomous agents capable of human-like social reasoning and interaction.

Acknowledgments

Suppressed for anonymity.

References

Abdelnabi, S.; Gomaa, A.; Sivaprasad, S.; Schönherr, L.; and Fritz, M. 2023. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games. *arXiv preprint arXiv:2309.17234*.

Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, 337–371. PMLR.

AL, A.; Ahn, A.; Becker, N.; Carroll, S.; Christie, N.; Cortes, M.; Demirci, A.; Du, M.; Li, F.; Luo, S.; et al. 2024. Project Sid: Many-agent simulations toward AI civilization. *arXiv preprint arXiv:2411.00114*.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 337–351.

Bara, C.-P.; CH-Wang, S.; and Chai, J. 2021. MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1112–1125.

Berg, J.; Dickhaut, J.; and McCabe, K. 1995. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1): 122–142.

Coletta, A.; Dwarakanath, K.; Liu, P.; Vyetrenko, S.; and Balch, T. 2024. LLM-driven Imitation of Subrational Behavior: Illusion or Reality? *arXiv preprint arXiv:2402.08755*.

Duan, J.; Zhang, R.; Diffenderfer, J.; Kailkhura, B.; Sun, L.; Stengel-Eskin, E.; Bansal, M.; Chen, T.; and Xu, K. 2024. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. *arXiv preprint arXiv:2402.12348*.

Engelbrecht, H. A.; and Schiele, G. 2013. Koekepan: Minecraft as a research platform. In *2013 12th Annual Workshop on Network and Systems Support for Games (NetGames)*, 1–3. IEEE.

Engelbrecht, H. A.; and Schiele, G. 2014. Transforming Minecraft into a research platform. In *2014 IEEE 11th Consumer Communications and Networking Conference (CCNC)*, 257–262. IEEE.

Fehr, E.; and Gächter, S. 2000. Fairness and retaliation: The economics of reciprocity. *Journal of economic perspectives*, 14(3): 159–182.

Fehr, E.; and Schmidt, K. M. 1999. A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3): 817–868.

Gürcan, Ö. 2024. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *HHAI 2024: Hybrid human AI systems for the social good*, 134–144.

Harsanyi, J. C. 1961. On the rationality postulates underlying the theory of cooperative games. *Journal of Conflict Resolution*, 5(2): 179–196.

Hong, S.; Zheng, X.; Chen, J.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Horton, J. J. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Hu, S.; Huang, T.; Ilhan, F.; Tekin, S.; Liu, G.; Kompella, R.; and Liu, L. 2024. A survey on large language model-based game agents. *arXiv preprint arXiv:2404.02039*.

Huang, J.-t.; Li, E. J.; Lam, M. H.; Liang, T.; Wang, W.; Yuan, Y.; Jiao, W.; Wang, X.; Tu, Z.; and Lyu, M. R. 2024. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. *arXiv preprint arXiv:2403.11807*.

Ke, L.; Tong, S.; Cheng, P.; and Peng, K. 2025. Exploring the frontiers of llms in psychological applications: A comprehensive review. *Artificial Intelligence Review*, 58(10): 305.

Kweon, S.; Nam, S.; Lim, H.; Hong, H.; and Choi, E. 2025. A Large-Scale Real-World Evaluation of LLM-Based Virtual Teaching Assistant. *arXiv preprint arXiv:2506.17363*.

Li, H.; Chen, J.; Yang, J.; Ai, Q.; Jia, W.; Liu, Y.; Lin, K.; Wu, Y.; Yuan, G.; Hu, Y.; et al. 2024. Legalagentbench: Evaluating llm agents in legal domain. *arXiv preprint arXiv:2412.17259*.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024a. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Lu, Y.; Aleta, A.; Du, C.; Shi, L.; and Moreno, Y. 2024b. Llms and generative agent-based models for complex systems research. *Physics of Life Reviews*, 51: 283–293.

Manning, B. S.; Zhu, K.; and Horton, J. J. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Mao, S.; Cai, Y.; Xia, Y.; Wu, W.; Wang, X.; Wang, F.; Ge, T.; and Wei, F. 2023. Alympics: Language agents meet game theory. *arXiv preprint arXiv:2311.03220*.

Mojang Synergies AB. 2022. Minecraft.

Nebel, S.; Schneider, S.; and Rey, G. D. 2016. Mining learning and crafting scientific experiments: a literature review on the use of minecraft in education and research. *Journal of Educational Technology & Society*, 19(2): 355–366.

O'Gara, A. 2023. Hoodwinked: Deception and cooperation in a text-based game for language models. *arXiv preprint arXiv:2308.01404*.

OpenAI. 2025. GPT-4o [Large language model]. Accessed: 2025-04-23.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Putniņš, T. J. 2012. MARKET MANIPULATION: A SURVEY. *Journal of Economic Surveys*, 26(5): 952–967.

Schmidgall, S.; Su, Y.; Wang, Z.; Sun, X.; Wu, J.; Yu, X.; Liu, J.; Liu, Z.; and Barsoum, E. 2025. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*.

Schmied, T.; Bornschein, J.; Grau-Moya, J.; Wulfmeier, M.; and Pascanu, R. 2025. LLMs are Greedy Agents: Effects of RL Fine-tuning on Decision-Making Abilities. *arXiv preprint arXiv:2504.16078*.

Simon, H. A. 1955. A behavioral model of rational choice. *The quarterly journal of economics*, 99–118.

Simon, H. A. 1997. *Models of bounded rationality: Empirically grounded economic reason*, volume 3. MIT press.

Sreedhar, K.; and Chilton, L. 2024. Simulating human strategic behavior: Comparing single and multi-agent llms. *arXiv preprint arXiv:2402.08189*.

Szymkiewicz, P. 2022. *Game engine for Minecraft Java Edition servers with Spigot API*. Ph.D. thesis, Zakład Projektowania Systemów CAD/CAM i Komputerowego Wspomagania Medycyny.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tversky, A.; Kahneman, D.; Slovic, P.; et al. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge.

Van Rooij, I.; Guest, O.; Adolfi, F.; de Haan, R.; Kolokolova, A.; and Rich, P. 2024. Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, 1–21.

Von Neumann, J.; and Morgenstern, O. 1947. Theory of games and economic behavior, 2nd rev.

Vrdoljak, J.; Boban, Z.; Vilović, M.; Kumrić, M.; and Božić, J. 2025. A review of large language models in medical education, clinical decision support, and healthcare administration. In *Healthcare*, volume 13, 603. MDPI.

Weber, M. 1978. *Economy and society: An outline of interpretive sociology*, volume 2. University of California press.

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.

Xu, C.; Guan, S.; Greene, D.; Kechadi, M.; et al. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291.

[letterpaper]article [submission]aaai2026 times helvet courier xcolor

Reproducibility Checklist

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) NA

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

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) yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously pub-

lished work) are accompanied by appropriate citations (yes/no/NA) NA

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

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