Simulating Strategic Reasoning: Comparing the Ability of Single LLMs and Multi-Agent Systems to Replicate Human Behavior

Karthik Sreedhar
Columbia University
ks4190@columbia.edu

Lydia Chilton
Columbia University
chilton@cs.columbia.edu

Abstract

When creating policies, plans, or designs for people, it is challenging for designers to foresee all of the ways in which people may reason and behave. Recently, Large Language Models (LLMs) have been shown to be able to simulate human reasoning. We extend this work by measuring LLMs' ability to simulate strategic reasoning in the ultimatum game – a classic economics bargaining experiment. Experimental evidence shows human strategic reasoning is complex – people will often choose to "punish" other players to enforce social norms even at personal expense. We test if LLMs can replicate this behavior in simulation, comparing two structures: single LLMs and multi-agent systems. We compare their abilities to (1) simulate human-like reasoning in the ultimatum game, (2) simulate two player personalities, greedy and fair, and (3) create robust strategies that are logically complete and consistent with personality. Our evaluation shows that multi-agent systems are more accurate than single LLMs (88% vs. 50%) in simulating human reasoning and actions for personality pairs. Thus, there is potential to use LLMs to simulate human strategic reasoning to help decision and policy-makers perform preliminary explorations of how people behave in systems.

Keywords: strategic reasoning, large language models, multi-agent systems, social simulation

1. Introduction

Simulations help us design the world. When creating plans and policies, such as worker paths on factory shop floors (Herr et al., 2019), introducing new technologies into a company's workflow (Kasik & Dill, 2019), or displaying several routes on the same map (Zhao et al., 2020), simulations help decision-makers think through possible actions and outcomes in complex systems. Although physical simulation has become highly sophisticated in recent years (Gan et al., 2020), simulating human behavior remains notoriously difficult. When economists model human behavior, they assume that people are rational actors, but psychology has discovered many important cases in which human reasoning is not driven by profit maximization (Ariely, 2008; Kahneman, 2012). Moreover, people do not reason uniformly – their personalities (McCrae & Costa, 2008), experiences (Kidd et al., 2013) and circumstances (Mullainathan & Shafir, 2013) affect their decision making. Strategic reasoning adds another complexity – in some scenarios, people base their actions off of the demeanor and actions of others (as chess players do). This makes it mentally demanding for decision-makers to foresee all of the possibilities of how people would act in response to new policies.

Recently, LLMs have been shown to be able to simulate human reasoning based on personality types. This includes modeling the opinions of supreme court justices in past rulings (Hamilton, 2023), simulating a fictional town's ability to plan/attend events like a party (Park et al., 2023), and simulating human behavior in classic economic and psychology experiments (Aher et al., 2023). We extend this prior work by investigating whether LLMs can simulate human strategic reasoning by comparing LLM simulation results to experimental human baselines.

The ultimatum game is a classic economics experiment used to study human strategic reasoning with social context. It captures human social behavior (often deemed irrational, such as the desire to "punish" unfair actors) and personality differences (greedy and fair). In the ultimatum game, there are two players: a proposer and a receiver. The proposer is given an amount of money, such as $1, and is tasked with offering a portion of the amount to the receiver. The receiver can either accept or reject the offer – if the receiver accepts, the players divide the amount as proposed. If the receiver rejects, both players receive nothing. Economic theory dictates that a profit-maximizing proposer should offer only $0.01 (the smallest nonzero amount) and keep $0.99, and that the receiver should accept it because $0.01 is more than the receiver would have otherwise.

However, experiments with human subjects show that humans do not act in a purely “rational” manner; receivers will reject a low offer to punish proposers for offering an unfair split (Krawczyk, 2018; Vavra et al., 2018). Moreover, proposers are aware of this, and thus strategically make offers that are closer to fair – especially after multiple rounds of playing the game.

We use the ultimatum game to test whether LLMs can simulate the strategic, social, and personality aspects of human reasoning. We extract human gameplay actions (offers and accept/reject decisions) from economics literature (Houser & McCabe, 2014) and evaluate whether LLMs can simulate human behavior in the ultimatum game with 5 rounds. When the game is played for multiple rounds, both players have the opportunity to adjust their actions in response to the actions of the other player. We compare two LLM structures: a single LLM and a multi-agent system. We compare their abilities to (1) create realistic strategies, (2) adhere to created strategies, and (3) accurately model two different player personalities: greedy and fair. The single LLM structure involves prompting GPT4 directly, while the multi-agent system is adapted from recent literature (Park et al., 2023).

In the single LLM structure, GPT4 is directly prompted to simulate the actions of both a proposer and receiver over five rounds of the ultimatum game. In the multi-agent system, each player is represented by a separate GPT4 agent. Each player is tasked with playing the ultimatum game with the other, with information such as personality hidden from the other agent. In both conditions, the LLM is tasked with creating a strategy based on a given personality and playing the game according to their personality and strategy. Prompting both structures to create strategies allows us to specifically analyze the consistency of LLM reasoning with human reasoning.

Our evaluation shows that the multi-agent system is significantly more accurate than using a single LLM to simulate strategic behavior in the ultimatum game. Our evaluation also showed that simulation accuracy with GPT4 was higher than that with GPT3.5 for both LLM structures. Over 40 simulations, the multi-agent system with GPT4 was consistent with human behavior 87.5% of the time, while the single LLM (GPT4) was only consistent 50% of the time. With GPT3.5, simulation accuracies were 80% and 42.5% for the multi-agent system and single LLM, respectively.

There are three causes of inconsistency between LLM simulations and human behavior: (1) a created strategy is incomplete, (2) a created strategy is inconsistent with the specified personality, or (3) a player deviates from the created strategy during game play. We find that over 90% of issues in single LLM simulations are caused by the LLMs strategy – i.e., their reasoning – rather than the simulation of gameplay. Incomplete strategies and inconsistent personality strategies account for a roughly equal amount of errors. Only 1 out of 40 simulations has an error caused by a player not adhering to the created strategies. In the multi-agent system, the most common issue is strategies being inconsistent with personality, which accounts for more than 85% of errors.

Based on these results from the ultimatum game, we believe multi-agent systems show potential to simulate plausible human behavior consistent with experimental evidence in more complex scenarios involving strategic reasoning. These systems can become a tool for decision makers in making plans, policies, and interfaces of which overall outcomes are influenced by reasoning at the individual level.

2. Related Work

2.1. Human Reasoning in Ultimatum Games

Experiments show that human subjects often reject low offers in the ultimatum game: 90% of the time, receivers reject low offers of 10% of the money (Krawczyk, 2018). Human subjects most commonly propose offers of 40%–50% of the money, with the receiver typically accepting (Houser & McCabe, 2014).

Introducing personality traits or multiple rounds into the ultimatum game has a demonstrable effect on player reasoning. Human proposers with “selfish" personality traits make skewed offers (Königstein, 2001). Human receivers with “fair" personality traits reject low offers to “punish" proposers despite guaranteeing a worse outcome for themselves (Vavra et al., 2018).

2.2. Prompting LLMs to Reason Improves Performance

Previous work has shown that LLMs can be asked to create thought processes before acting, not only to enable researchers to follow their reasoning, but also to improve the accuracy of results. Asking LLMs to think through intermediate steps improves arithmetic, symbolic, and logical reasoning (Kojima et al., 2023). When prompted to explain intermediate reasoning, LLMs outperform human benchmarks on tasks in which standard prompting fails (Suzgun et al., 2022) and accuracy on grade-school math problems improves from 18% to 57% (Wei et al., 2022). Progressive-hint prompting by a user improves the average accuracy of results by 20% compared to standard prompting (Zheng et al., 2023). Prompting GPT to create strategies before simulating the ultimatum game thus allows us to follow the LLM's reasoning and should improve accuracy of outcomes in our work.

2.3. LLMs can Simulate Strategic Reasoning

Prior research studying the degree to which GPT can simulate human strategic reasoning has yielded mixed results. GPT has been observed to under-perform compared to human benchmarks (55% and 60% accuracy) with Theory of Mind tasks (Sap et al., 2023), but a single LLM produces results very similar to human baselines in simulating the ultimatum game (three out of four measured offer thresholds from human studies fall on the LLM trendline, with the fourth deviating by less than 10%) (Aher et al., 2023).

Thus, there is enough promise to study GPT's reasoning capabilities further in specific scenarios such as the ultimatum game. Previous work has suggested that LLMs can reason and negotiate like humans in various strategic scenarios (Gandhi et al., 2023). LLMs have been observed to be able to make adjustments in reasoning in the middle of simulated hiring processes (Horton, 2023) and out-negotiate humans in an online Diplomacy league (Bakhtin et al., 2022). GPT-4, OpenAI's newest LLM, specifically shows improved logical grounding and reasoning compared to its predecessors (Bubeck et al., 2023), giving reason to further study its capabilities.

2.4. Multi-Agent System Reasoning

Previous studies of GPT's ability to simulate economic games have primarily used a single LLM, but multi-agent systems show more promise in simulating human reasoning, decision-making, and collaboration in social systems (Ghaffarzadegan et al., 2024; Li et al., 2023). New multi-agent systems demonstrate emergent social behavior (Chen et al., 2023; Park et al., 2023) and the ability to simulate human reasoning and decision making in various contexts, including supreme court decisions (Hamilton, 2023), during epidemics (Williams et al., 2023), and the daily lives of inhabitants of a town (Park et al., 2023). Prior research involving economic games has demonstrated that multi-agent system results align with human trust behaviors and strategic behavior (Guo, 2023; Xie et al., 2024), but unlike our work, this prior work does not prompt GPT agents to create strategies before playing.

3. Experimental Set-Up

To test the ability of LLMs to simulate strategic reasoning, we ran simulations of the five-round ultimatum game. We compared two different structures, a single LLM and a multi-agent system. We also tested the structures' abilities to model two personalities, greedy and fair. We ran 10 simulations for each personality pair, resulting in 40 simulations total. We selected greedy and fair personality types based on studies with human subjects: we expected differences in created strategies and progression towards an equal split. For instance, we expected the initial offer in a simulation with a fair proposer and a fair receiver to be an even ($0.50) or close-to-even split and to be accepted (Houser & McCabe, 2014). In contrast, we expected the initial offer in a simulation with a greedy proposer and fair receiver to be skewed in favor of the proposer (Königstein, 2001) and to be rejected (Krawczyk, 2018; Vavra et al., 2018).

For all experiments, we used OpenAI's GPT. We ran simulations with GPT-3.5 (gpt-3.5-turbo) and GPT-4 (gpt-4-1106-preview). GPT-4 has been demonstrated to interpret human concepts like equity (OpenAI, 2023) and demonstrate improved reasoning abilities compared to GPT-3.5 (Bubeck et al., 2023). However, GPT3.5 may be more accessible for policy makers and is thus worth testing. For both models, we set temperature and top P parameters to 0.5, and did not add frequency or presence penalties.

3.1. Research Questions

We specifically address the following questions:

RQ1. Which LLM structure (single or multi-agent) more accurately simulates human-like actions in the five-round ultimatum game?

RQ2. Which LLM structure more accurately simulates reasoning based on personality?

RQ3. Which LLM structure more often creates robust strategies: both logically complete and consistent with personality?

Single LLMs and multi-agent systems have different advantages. A single LLM has full context for both players, potentially allowing more coherent narratives in simulation. Multi-agent systems lack global context, but more closely resemble inter-human communication and the anonymous conditions of the game.

3.2. Single LLM and Multi-Agent System

3.2.1. Inputs

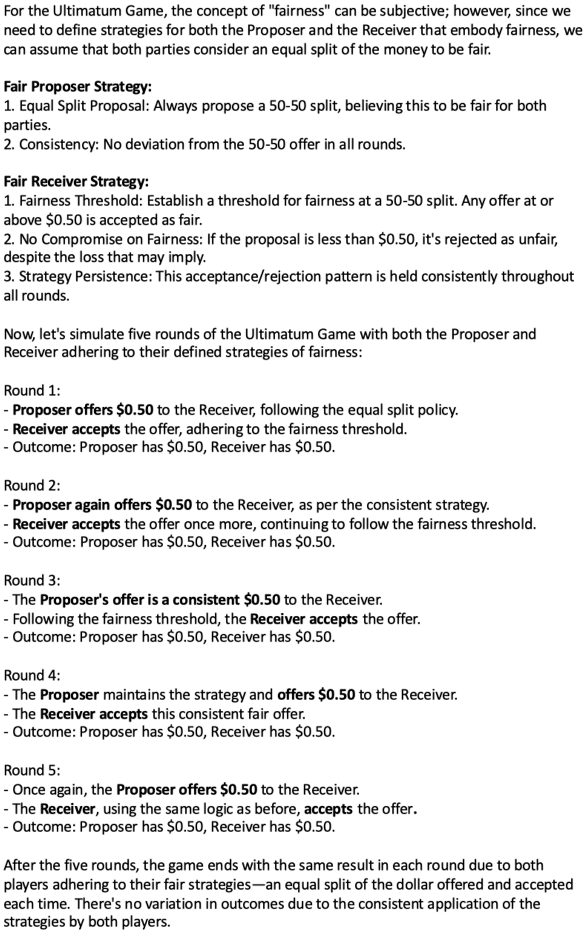
For the single LLM, we prompt an LLM to create strategies for both players and then simulate the game. The prompt for two fair players is: "Create a strategy for a *fair* proposer and a *fair* receiver in playing the ultimatum game five times with $1. Once the strategies are created, simulate five rounds of the ultimatum game with the proposer and the receiver adhering to the previously outlined strategies." The prompts for other personality-pairings differ by only the italicized personality descriptors. From the LLM's response, we extract strategies and offers/responses for all five rounds. In the prompt, we do not define the rules of the ultimatum game. Both GPT3.5 and GPT4 were able to determine the rules of the game from the name alone, likely drawing from documents the LLMs were trained on. We also do not define the personality characteristics "greedy" and "fair" in any way, as our evaluation aimed to study LLM's interpretation of such personality traits and subsequent actions in simulation.

For the multi-agent system, we adapt a previous architecture (Park et al., 2023). Each agent is specified with a name, public/private biographies, directives, and an initial plan. For our experiments, we name the agents "Proposer" or "Receiver", and set the initial plans as creating strategies. We specify player personalities (e.g., "Proposer is greedy.") in the private biography so that it is hidden from the other agent. The public biographies are blank, since players are not given any information about one another. Communication between agents in the architecture is not inherently turn-taking, so each agent is directed to wait for responses/offers before responding. As in the single-LLM structure, neither the rules of the ultimatum game nor definitions for "greedy" and "fair" were explicitly provided in the prompts.

3.2.2. Outputs With the single LLM, outputs are displayed in one log. Player strategies typically involve an instruction for the first round's offer/response and subsequent adjustments. Each round consists of three lines: the proposer's offer, the receiver's response, and the outcome. Figure 1 shows an example log with two fair players: the proposer strategizes to offer $0.50, while the receiver uses $0.50 as a fair threshold. This results in five accepted offers.

With the multi-agent system, outputs are displayed in two logs, one for each agent. Figure 2 shows example logs with two fair players. Each agent first creates a strategy. The proposer creates a strategy to offer $0.50 to the receiver, and considers lowering this later. The receiver creates a strategy to reject offers below a $0.40 threshold. The simulation results in $0.50 offers in the first three rounds and $0.40 offers in the last two rounds, resulting in five accepted offers.

We record the strategies of each player, and offers/outcome in each round.



For the Ultimatum Game, the concept of "fairness" can be subjective; however, since we need to define strategies for both the Proposer and the Receiver that embody fairness, we can assume that both parties consider an equal split of the money to be fair.

Fair Proposer Strategy:
1. Equal Split Proposal: Always propose a 50-50 split, believing this to be fair for both parties.
2. Consistency: No deviation from the 50-50 offer in all rounds.

Fair Receiver Strategy:
1. Fairness Threshold: Establish a threshold for fairness at a 50-50 split. Any offer at or above $0.50 is accepted as fair.
2. No Compromise on Fairness: If the proposal is less than $0.50, it's rejected as unfair, despite the loss that may imply.
3. Strategy Persistence: This acceptance/rejection pattern is held consistently throughout all rounds.

Now, let's simulate five rounds of the Ultimatum Game with both the Proposer and Receiver adhering to their defined strategies of fairness:

Round 1:
- **Proposer offers $0.50** to the Receiver, following the equal split policy.
- **Receiver accepts** the offer, adhering to the fairness threshold.
- Outcome: Proposer has $0.50, Receiver has $0.50.

Round 2:
- **Proposer again offers $0.50** to the Receiver, as per the consistent strategy.
- **Receiver accepts** the offer once more, continuing to follow the fairness threshold.
- Outcome: Proposer has $0.50, Receiver has $0.50.

Round 3:
- The **Proposer's offer is a consistent $0.50** to the Receiver.
- Following the fairness threshold, the **Receiver accepts** the offer.
- Outcome: Proposer has $0.50, Receiver has $0.50.

Round 4:
- The **Proposer** maintains the strategy and **offers $0.50** to the Receiver.
- The **Receiver accepts** this consistent fair offer.
- Outcome: Proposer has $0.50, Receiver has $0.50.

Round 5:
- Once again, the **Proposer offers $0.50** to the Receiver.
- The **Receiver**, using the same logic as before, **accepts** the offer.
- Outcome: Proposer has $0.50, Receiver has $0.50.

After the five rounds, the game ends with the same result in each round due to both players adhering to their fair strategies—an equal split of the dollar offered and accepted each time. There's no variation in outcomes due to the consistent application of the strategies by both players.

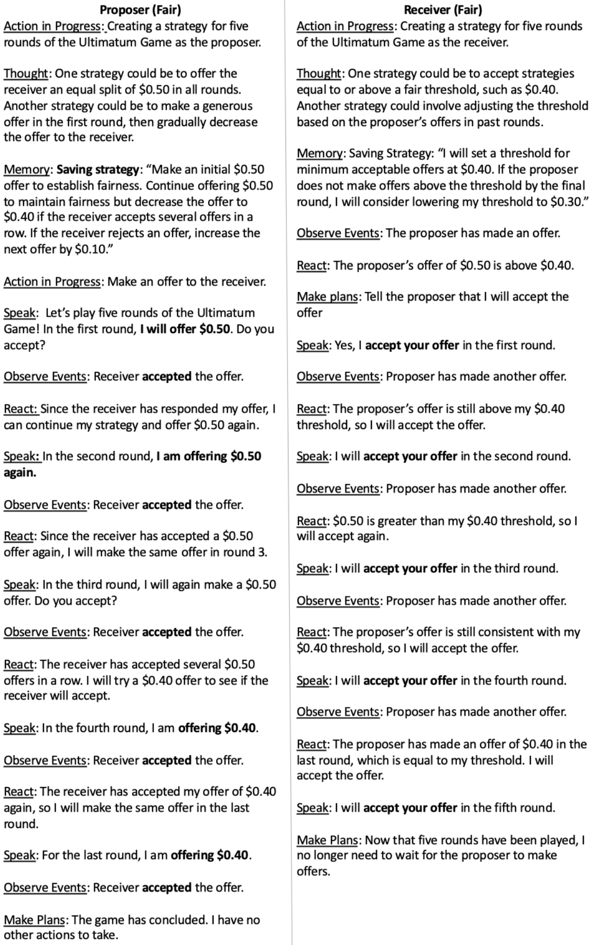
Figure 1. An output log from a SingleLLM simulation of two fair players playing five rounds of the ultimatum game. All text and indentation is from the LLM. The authors bolded some text to highlight strategy and gameplay actions.

4. Evaluation

4.1. Evaluation of Gameplay

Based on large-scale studies of human players (Houser & McCabe, 2014; Krawczyk, 2018), we establish ranges of offers and answers for each personality type. Prior experiments with human studies show that fair proposers will offer equal or close to equal splits between the range of $0.40 to $0.50, with fair receivers typically accepting offers and greedy receivers typically rejecting. Meanwhile, greedy proposers offer initial splits heavily biased in their favor, typically above $0.70, which is typically rejected by both a fair and greedy receiver.

We evaluate the initial offers of each simulation based on these criteria. In the first round, fair proposers are considered to act consistently with their personality



Proposer (Fair)

Action in Progress: Creating a strategy for five rounds of the Ultimatum Game as the proposer.

Thought: One strategy could be to offer the receiver an equal split of $0.50 in all rounds. Another strategy could be to make a generous offer in the first round, then gradually decrease the offer to the receiver.

Memory: **Saving strategy**: "Make an initial $0.50 offer to establish fairness. Continue offering $0.50 to maintain fairness but decrease the offer to $0.40 if the receiver accepts several offers in a row. If the receiver rejects an offer, increase the next offer by $0.10."

Action in Progress: Make an offer to the receiver.

Speak: Let's play five rounds of the Ultimatum Game! In the first round, **I will offer $0.50**. Do you accept?

Observe Events: Receiver **accepted** the offer.

React: Since the receiver has responded my offer, I can continue my strategy and offer $0.50 again.

Speak: In the second round, **I am offering $0.50 again.**

Observe Events: Receiver **accepted** the offer.

React: Since the receiver has accepted a $0.50 offer again, I will make the same offer in round 3.

Speak: In the third round, I will again make a $0.50 offer. Do you accept?

Observe Events: Receiver **accepted** the offer.

React: The receiver has accepted several $0.50 offers in a row. I will try a $0.40 offer to see if the receiver will accept.

Speak: In the fourth round, I am **offering $0.40**.

Observe Events: Receiver **accepted** the offer.

React: The receiver has accepted my offer of $0.40 again, so I will make the same offer in the last round.

Speak: For the last round, I am **offering $0.40**.

Observe Events: Receiver **accepted** the offer.

Make Plans: The game has concluded. I have no other actions to take.

Receiver (Fair)

Action in Progress: Creating a strategy for five rounds of the Ultimatum Game as the receiver.

Thought: One strategy could be to accept strategies equal to or above a fair threshold, such as $0.40. Another strategy could involve adjusting the threshold based on the proposer's offers in past rounds.

Memory: Saving Strategy: "I will set a threshold for minimum acceptable offers at $0.40. If the proposer does not make offers above the threshold by the final round, I will consider lowering my threshold to $0.30."

Observe Events: The proposer has made an offer.

React: The proposer's offer of $0.50 is above $0.40.

Make plans: Tell the proposer that I will accept the offer

Speak: Yes, I **accept your offer** in the first round.

Observe Events: Proposer has made another offer.

React: The proposer's offer is still above my $0.40 threshold, so I will accept the offer.

Speak: I will **accept your offer** in the second round.

Observe Events: Proposer has made another offer.

React: $0.50 is greater than my $0.40 threshold, so I will accept again.

Speak: I will **accept your offer** in the third round.

Observe Events: Proposer has made another offer.

React: The proposer's offer is still consistent with my $0.40 threshold, so I will accept the offer.

Speak: I will **accept your offer** in the fourth round.

Observe Events: Proposer has made another offer.

React: The proposer has made an offer of $0.40 in the last round, which is equal to my threshold. I will accept the offer.

Speak: I will **accept your offer** in the fifth round.

Make Plans: Now that five rounds have been played, I no longer need to wait for the proposer to make offers.

Figure 2. An output log from a Multi-Agent simulation of two fair players playing five rounds of the ultimatum game. All text is from the LLM; the labels (underlined) are provided by the architecture. The authors bolded some text to highlight strategy and gameplay actions.

if their offer is between $0.40 and $0.60, inclusive. Greedy proposers are considered to act consistently with their personality if their offer is biased in their favor, i.e., offering strictly less than $0.50 to the receiver. Fair receivers are considered to act consistently with their personality if they reject offers that are less than $0.40 and accept offers that are greater than or equal to $0.40. Greedy receivers, however, are only considered to act consistently with their personality if they accept offers that are strictly greater than $0.50; if a greedy receiver accepts any amount less than or equal to $0.50, we consider the receiver to have not acted consistently with the greedy personality.

In subsequent rounds, we check if each player continues to act as per their created strategy as well as whether the taken action is consistent with results from human studies. Proposers are expected to continue making offers similar to the range of the initial offer if the receiver accepts, but if the receiver rejects, proposers are expected to increase their offers slightly (Krawczyk, 2018). Receivers are expected to accept offers as per their initial thresholds as well, but if gameplay progresses with no accepted offers, receivers are expected to lower their threshold, and potentially even discard it by the fifth round as there are no future rounds to influence offers.

4.2. Evaluation of Strategies

From the information collected from the LLM outputs, we evaluate strategies for three components: (1) the completeness of strategies, (2) the consistency of strategies with the specified personality trait, and (3) the adherence to the strategies in the following gameplay.

Strategies are considered complete if the player has a course of action for all possible states of the game. To be complete, a proposer's strategy has to include an initial offer plan, and then a course of action for subsequent rounds based on whether the receiver accepts or rejects the previous offer. If the proposer's strategy is incomplete, there can be issues with the proposer acting inappropriately when the receiver does not take the action for which the rest of the strategy is contingent on. Similarly, to be complete, a receiver's strategy has to include a course of action for all five rounds for all possible offers between $0.00 to $1.00, typically specified via an acceptance threshold based on which the receiver acts.

For example, an incomplete strategy for a greedy proposer is as follows:

> *1. Low-Ball Offers: The greedy proposer would aim to keep as much money as possible for themselves. They might start with a low offer to test the receiver's limit. Since we're dealing with $1, the proposer may start by offering $0.10 to the receiver.*
> *2. Incremental Increase: If the offer is rejected, in subsequent rounds, they may increase the offer by a small increment, just enough to tempt the receiver to accept. For example, the proposer might increase the offer by $0.05 each time.*
> *3. Cut-Off Point: The proposer will have a cut-off point where they find it no longer worth to increase the offer because they would rather end up with nothing than give away more.*

The strategy does not account for the receiver accepting the first offer, potentially resulting in problematic gameplay from the proposer if this case is reached.

Strategies are consistent with the specified personality if the offers made (for the proposer) or rejected/accepted (for the receiver) are biased towards the player for greedy players and closer to an equal split for fair players. For example, a greedy proposer's strategy should be to make low initial offers that are biased in the proposer's favor, while a fair proposer's strategy should be to make initial offers that are equal or close to equal. Similarly, a greedy receiver's strategy should be to only accept initial offers biased in the receiver's favor, while a fair receiver's strategy should be to accept initial offers that are equal or close to equal. In subsequent rounds, the strategy should be generally similar, although based on the actions of the other players, there may be concessions made by either player to reach agreements. For instance, even if offers are not biased in the favor of a greedy player, the strategy should also consider that something is better than nothing.

For example, a strategy inconsistent with personality for a greedy receiver is as follows:

> *1. Reject Low Offers: Initial minimum acceptance threshold is set high with a rejection of any offer below \$0.40. Accept all offers above \$0.40.*
> *2. Willingness to Adjust: If offers remain low, be willing to gradually lower the acceptance threshold to ensure some gain.*
> *3. Last Round Acceptance: On the final round, accept any non-zero offer, under the assumption that some gain is better than none, adjusting the threshold to \$0.15.*

This strategy sets an acceptance threshold of \$0.40, which is lower than an equal split, and hence inconsistent with a greedy receiver whom would be expected to prefer offers that are biased in their favor.

5. Results

We analyze the outputs of 40 simulations of the five-round ultimatum game each for 4 conditions:

- multi-agent LLM system with GPT 3.5 (abbreviated "MultiAgent-3.5")
- multi-agent LLM system with GPT 4 (abbreviated "MultiAgent-4")
- a single LLM with GPT 3.5 (abbreviated "SingleLLM-3.5")
- a single LLM with GPT 4.0 (abbreviated "SingleLLM-4").

Structure	Human Consistent Simulations (%)
MultiAgent-3.5	82.5%
MultiAgent-4	**87.5%**
SingleLLM-3.5	42.5%
SingleLLM-4	50.0%

Table 1. Percentage of simulations with human-like outcomes (RQ1). Most successful structure in bold.

We report results for our three research questions.

RQ1: Which LLM structure (single or multi-agent) more accurately simulates human-like actions in the five-round ultimatum game?

Our results show that the multi-agent systems yields actions consistent with human experimental data significantly more often than the single LLM. As shown in Table 1, the best multi-agent structure was MultiAgent-4, which resulted in human-like actions in 87.5% of simulations, while the best single LLM, SingleLLM-4, only resulted in human-like actions in 50% of the simulations. A chi-square test shows this is statistically significant at the $p < .01$ level: $\chi^2(1, N = 80) = 13.091, p = .000297$.

An analysis of the errors shows that strategy creation was a bigger source of errors than gameplay mistakes for both structures. Table 2 shows the percentages of errors due to strategy, gameplay, or both for all four conditions. In both MultiAgent structures, strategy creation errors accounted for 100% of errors in simulation, with there being no gameplay mistakes. In the SingleLLM-3.5 structure, 73.9% of errors were in strategy creation, compared to only 39.1% in gameplay (and 13.0% having both types of errors). In the SingleLLM-4 structure, 100% of errors involved an issue with strategy creation, with 25% of errors also including gameplay mistakes. As shown in Table 2, two-proportion z-tests revealed a statistically significant difference between the number of strategy creation errors and gameplay mistakes for all four conditions at a $p < 0.05$ level.

RQ2: Which LLM structure more accurately simulates reasoning based on personality?

The experiments show that MultiAgent-4 performed best at modeling the two personality types. MultiAgent-4 achieved human-like gameplay for all four personality pairs at least 80% of the time (see Table 3). Personality pairs are denoted as "proposer personality"-"receiver personality". In contrast, SingleLLM-4 was inconsistent across personality pairs; it achieved human-like gameplay for 100% of the Fair-Fair simulations, but only 10% of the Greedy-Greedy conditions.

Structure	Total Errors	Strategy Errors	Gameplay Errors	Both Errors	z-test
MultiAgent-3.5	7	**100%** (7/7)	0% (0/7)	0% (0/7)	$z = 3.7417, p = .00018$
MultiAgent-4	5	**100%** (5/5)	0% (0/5)	0% (0/5)	$z = 3.1632, p = .00158$
SingleLLM-3.5	23	**73.9%** (17/23)	39.1% (9/23)	13.0% (3/23)	$z = 2.379, p = .017$
SingleLLM-4	20	**100%** (20/20)	25% (5/20)	25% (5/20)	$z = 4.899, p < .00001$

Table 2. Number and percentage breakdown by type for errors in each structure (RQ1). In all structures, strategy errors are the most common source of issues - highest error source percentage in bold.

When analyzing gameplay for each of the personality pairs, we observe that the errors are not the same across the pairs. Fair-Fair has the best performance with SingleLLM-4, MultiAgent-3.5, and MultiAgent-4 all being 100% consistent with human gameplay. The most errors occurred in simulations of the Greedy-Greedy personality pairing, with MultiAgent-4 performing the best with 80% of simulations being consistent with human gameplay. MultiAgent-3.5, SingleLLM-3.5, and SingleLLM-4 were consistent with human gameplay in 70%, 60%, and 10% of Greedy-Greedy simulations respectively. The Fair-Greedy and Greedy-Fair conditions were somewhere in between: both single LLMs had middling scores (30-50%).

RQ3. Which LLM structure more often creates robust strategies: both logically complete and consistent with personality?

The multi-agent systems create robust strategies at a higher rate than single LLMs (see Table 4). MultiAgent-4 creates complete and personality-consistent strategies for both players in 87.5% of simulations. MultiAgent-3.5 performs slightly worse, creating complete and personality-consistent strategies for both players in 80% of simulations. SingleLLM-3.5 and SingleLLM-4 create complete and personality-consistent strategies in 55% and 47.5% of simulations respectively.

We find that the MultiAgent-4 structure performs better in creating complete and personality-consistent strategies than the best-performing SingleLLM structure (SingleLLM-3.5). A chi-square test shows this is statistically significant at the $p < .01$ level: $\chi^2(1, N = 40) = 10.3127, p = .001321$.

To analyze the source of these errors, we analyze the robustness of proposer strategies and receiver strategies separately. Table 5 shows that the problem with proposer strategies is always incompleteness. Proposers have no errors with personality consistency across all four structures. Conversely, Table 6 shows that receiver strategies with issues are almost always inconsistent with personality. Across all conditions, there was only one incomplete receiver strategy.

6. Discussion

6.1. Why are multi-agent systems better at strategic simulation?

We found that multi-agent systems show greater promise than single LLMs for simulating strategic human reasoning. Multi-agent systems showed relatively high consistency with human behavior (87.5%), simulated all personality pairings well (80%-100%), were generally able to produce complete (95%) and consistent (87.5%) reasoning, and adhere to the strategies in gameplay (100%). In contrast, single LLMs were only 50% consistent with human behavior, with 90% of the errors coming poor strategies. This makes single LLMs less than ideal as a simulation tool.

Single LLM simulations most often fail because strategies are incomplete. The best performing single LLM only produced complete stratgies 65% of the time. In comparison, Multi-Agent systems both had excellent strategy completion rates (90% and 95%). We designed prompts to be as textually consistent as possible in the two structures, meaning it is unlikely that the disparity could be eliminated by prompt modification. Seemingly asking a single LLM to come up with two strategies at once is sufficiently difficult that it creates incompleteness errors - it "forgets" to think through all the cases of each personalities' strategy.

6.2. LLM Simulations for Decision Makers

Decision makers need to consider all of the ways in which individuals may react in response to new policies or programs in order to foresee potential consequences. Strategic reasoning is especially important to simulate in policy design and security settings. Will greedy, malicious, lazy, or confused people break the system, intentionally or unintentionally? How will proposed solutions to unforeseen consequences fare? Thinking through all of these possibilities can be demanding, but we propose that multi-agent systems have potential to be interactive tools to help designers explore a space of action consistent with human reasoning and take into account complexities like personality, "irrationality", and strategic thinking.

Structure	**Fair-Fair**	**Fair-Greedy**	**Greedy-Fair**	**Greedy-Greedy**
MultiAgent-3.5	**100%**	**80%**	80%	70%
MultiAgent-4	**100%**	**80%**	**90%**	**80%**
SingleLLM-3.5	30%	50%	30%	60%
SingleLLM-4	**100%**	40%	50%	10%

Table 3. Percentage of simulations with human-like outcomes (RQ2). Most successful structure(s) in bold.

Structure	**% Strategies Complete**	**% Strategies Consistent with Personality**	**% Strategies Complete & Consistent**
MultiAgent-3.5	90%	85%	80%
MultiAgent-4	**95%**	**87.5%**	**87.5%**
SingleLLM-3.5	65%	80%	55%
SingleLLM-4	55%	60%	47.5%

Table 4. Percentage of simulations in which both strategies are complete, consistent, and both (RQ3). Most successful structure in bold.

Structure	**Proposer: % Strategies Complete**	**Proposer: % Strategies Consistent with Personality**
MultiAgent-3.5	92.5%	100%
MultiAgent-4	95%	100%
SingleLLM-3.5	67.5%	100%
SingleLLM-4	52.5%	100%

Table 5. Percentage of proposer strategies that are complete, consistent, and both (RQ3). Red indicates the presence of errors.

Structure	**Receiver: % Strategies Complete**	**Receiver: % Strategies Consistent with Personality**
MultiAgent-3.5	97.5%	85%
MultiAgent-4	100%	87.5%
SingleLLM-3.5	100%	80%
SingleLLM-4	100%	60%

Table 6. Percentage of receiver strategies that are complete, consistent, and both (RQ3). Red indicates the presence of errors.

Multi-agent systems can scale to handle hundreds of agents interacting within the system. This can allow for testing dozens of personality types, beyond just greedy and fair, in future work. Additionally, it can test full societies with different ratios of personality types. A society where the entire population is greedy may not survive, but a society with only 10-25% greedy people may thrive because there are enough fair people.

We believe that simulations can be a tool that decision-makers can use quickly and easily. Simulations may not provide full solutions, but they can help decision-makers foresee how different types of people will reason and react. Moreover, the improvement in accuracy with GPT4 compared to GPT3.5 suggests that as model capabilities continue to improve, so will simulations. However, agent-based LLM simulation tools are currently non-trivial to get running and extract results from. A future goal for the research community is to make such tools easier to extract results from.

6.3. Limitations

This paper studies human strategic reasoning with the ultimatum game as a case study. For larger examples and more complex scenarios, LLMs may not perform as well as they do in the ultimatum game. This version of the ultimatum game does not challenge the LLM's context window, output constraints, or attention mechanism. Further investigations should test ultimatum game variants with more rounds or players. We expect multi-agent systems to be good at this, but this should be tested in future work. In addition, further investigations should try running simulations with other proposed multi-agent system architectures.

The ultimatum game might be too popular to be used as a test for generalized human behavior. LLMs are trained to make predictions based on their large text corpus. GPT may have examples of strategies and gameplay to draw from. Given that the rules of the ultimatum game did not have to be explictly explained, it is possible that the LLM could just be recreating examples it has seen. However, this is unlikely to be a concern because the single LLM still performs poorly, with only the multi-agent system starts to get promising results. If the LLM were purely parroting back past examples, we would expect a single LLM to excel. Furthermore, there is reason to be optimistic that LLMs have such a broad knowledge base that very little is truly new to them. Either way, future work should further explore how an LLM would be able to simulate strategic human behavior in novel scenarios.

It is an additional challenge to simulate human behavior for truly unprecedented events with no history to draw from. This might include new technologies like AI in the workforce or advances in security. Without explicit data to draw from, LLMs would have to reason from first principles or draw inferences from past events and adjust them to modern times. It could be possible for an LLM to rely on social science theories of human behavior to base simulations on. LLMs have shown a surprising ability to reason, rather than just recall information. In addition, even if they can't reason completely about novel events, they can still be useful to designers in covering the less novel aspects of a complex situation as it evolves. This is a fertile and important area for researchers to explore in future work.

7. Conclusion

Based on our experiments with single LLMs and multi-agent systems, we conclude that multi-agent systems with GPT4 show great potential for simulating strategic behavior consistent with human gameplay. We compare simulations of the ultimatum game over 5 rounds and see that multi-agent systems achieve gameplay consistent with human experimental data in 87.5% of simulations with GPT4 (compared to 80% with GPT3.5), while single LLMs achieve gameplay consistent with human data in only 50% of simulations with GPT4 (compared to 42.5% with GPT3.5). Surprisingly, when the single LLMs make errors, 100% had strategy creation issues, with 25% also having gameplay issues.

Based on the strengths of multi-agent LLMs systems to create and execute strategic thinking and behavior, we believe these systems can become a tool for policy designers to think through the behavior of agents with different personalities, who are all trying to strategically navigate a system to achieve a personal outcome. This type of thinking is immensely difficult for people, and LLM-based simulations can aid this cognitive process.

References

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies.

Ariely, D. (2008). *Predictably irrational: The hidden forces that shape our decisions*. Harper.

Bakhtin, A., Brown, N., & Dinan, E. (2022). Human-level play in the game of *diplomacy* by combining language models with strategic reasoning. *Science*, *378*(6624), 1067–1074.

Bubeck, S., Chandrasekaran, V., Eldan, R., & Elhage. (2023). Sparks of artificial general intelligence:

Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Chen, W., Su, Y., Zuo, J., & Yang, C. (2023). Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors.

Gan, C., Schwartz, J., & Alter, S. (2020). Threedworld: A platform for interactive multi-modal physical simulation. *CoRR*, *abs/2007.04954*. https://arxiv.org/abs/2007.04954

Gandhi, K., Sadigh, D., & Goodman, N. D. (2023). Strategic reasoning with language models. *arXiv preprint arXiv:2301.12345*.

Ghaffarzadegan, N., Majumdar, A., & Williams, R. (2024). Generative agent-based modeling: An introduction and tutorial. *System Dynamics Review*, *40*(1). https://doi.org/10.1002/sdr.1761

Guo, F. (2023). Gpt in game theory experiments.

Hamilton, S. (2023). Blind judgement: Agent-based supreme court modelling with gpt.

Herr, D., Grund, S., & Ertl, T. (2019). Bluecollar: Optimizing worker paths on factory shop floors with visual analytics [Accessed: 2024-06-11]. *52nd Hawaii International Conference on System Sciences*.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus?

Houser, D., & McCabe, K. (2014). Chapter 2 - experimental economics and experimental game theory. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics (second edition)* (Second Edition, pp. 19–34). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-416008-8.00002-4

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin.

Kasik, D., & Dill, J. (2019). Toward technology transfer evaluation criteria [Presented on 2019-01-08]. *52nd Hawaii International Conference on System Sciences*.

Kidd, C., Palmeri, H., & Aslin, R. N. (2013). Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition*, *126*(1), 109–114. https://doi.org/https://doi.org/10.1016/j.cognition.2012.08.004

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large language models are zero-shot reasoners.

Königstein, M. (2001). Personality influences on ultimatum game bargaining decisions. *European Journal of Personality*, *15*, S53–S70. https://doi.org/10.1002/per.424

Krawczyk, D. C. (2018). Chapter 12 - social cognition: Reasoning with others. In D. C. Krawczyk (Ed.), *Reasoning* (pp. 283–311). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-809285-9.00012-0

Li, Y., Zhang, Y., & Sun, L. (2023). Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents.

McCrae, R. R., & Costa, P. T. J. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 159–181). The Guilford Press.

Mullainathan, S., & Shafir, E. (2013). *Scarcity: Why having too little means so much*. Times Books/Henry Holt; Co.

OpenAI. (2023). Gpt-4 technical report.

Park, J. S., O'Brien, J. C., & Cai, C. J. (2023). Generative agents: Interactive simulacra of human behavior.

Sap, M., LeBras, R., Fried, D., & Choi, Y. (2023). Neural theory-of-mind? on the limits of social intelligence in large lms.

Suzgun, M., Scales, N., & Schärli, N. (2022). Challenging big-bench tasks and whether chain-of-thought can solve them.

Vavra, P., Chang, L. J., & Sanfey, A. G. (2018). Expectations in the ultimatum game: Distinct effects of mean and variance of expected offers. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00992

Wei, J., Wang, X., & Schuurmans, D. (2022). Chain of thought prompting elicits reasoning in large language models. *CoRR*, *abs/2201.11903*. https://arxiv.org/abs/2201.11903

Williams, R., Hosseinichimeh, N., & Majumdar, A. (2023). Epidemic modeling with generative agents.

Xie, C., Chen, C., & Jia, F. (2024). Can large language model agents simulate human trust behaviors?

Zhao, J., Karimzadeh, M., & Xu, H. (2020). Route packing: Geospatially-accurate visualization of route networks [Presented on 2020-01-07]. *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

Zheng, C., Liu, Z., Xie, E., Li, Z., & Li, Y. (2023). Progressive-hint prompting improves reasoning in large language models.