# Your Title Here [⋆]

Yifan Li[1]

Duke Kunshan University, Kunshan, Jiangsu 215316, China
yl772@duke.edu
LinkedIn

**Abstract.** This paper explores the integration of Large Language Models (LLMs) into the framework of trust games, focusing on the use of AI-driven fact-checking mechanisms to establish trust levels among participants. Trust games, fundamental in understanding cooperative behavior in economics, typically involve scenarios where players decide whether to cooperate based on the expectation of reciprocal actions. However, information asymmetry can significantly skew perceived risks and rewards. By implementing fact-checking mechanisms through LLMs, this study aims to quantify trust dynamically and reduce the uncertainty inherent in these decisions. The primary contribution of this research lies in its potential to enhance the accuracy of trust assessments in economic interactions, thereby promoting more efficient and transparent markets. *Notes: In submission to Problem Set 2 for COMPSCI/ECON 206 Computational Microeconomics, 2024 Spring Term (Seven Week - Fourth) instructed by Prof. Luyao Zhang at Duke Kunshan University.*

**Keywords:** computational economics · game theory · innovative education · provide more keywords here

## 1 Introduction

The concept of trust plays a pivotal role in the dynamics of economic interactions, influencing decisions in markets, negotiations, and cooperative ventures. Traditional game theory often assumes rational agents with perfect information, an assumption that rarely holds true in real-world scenarios. The emergence of digital platforms and online marketplaces has further complicated the dynamics of trust due to increased anonymity and potential information asymmetry.

Large Language Models (LLMs), powered by advanced machine learning algorithms, have shown promise in processing vast amounts of unstructured data

[⋆] **Acknowledgments**: when writing acknowledgments for a research paper, start by expressing sincere gratitude to your instructor and supervisor for their guidance. Then, extend thanks to your colleagues for their support, and to your classmates for their engaging discussions and input. Conclude with a personal note of appreciation for friends and family who offered encouragement throughout your research endeavor. This sequence ensures that gratitude is shown to both professional and personal supporters who have contributed to your academic journey.

to extract meaningful insights. This research proposes the novel integration of LLMs' fact-checking capabilities into trust games[1], aiming to assess the truthfulness of participants' historical data and statements to dynamically assign trust levels. Such an approach could significantly impact economic theories and practices by providing a more robust foundation for trust assessment, ultimately leading to more effective and fair economic outcomes.

This paper is structured as follows: Section 2 reviews the literature on trust games and the role of information in economic decisions. Section 3 describes the methodology for integrating LLMs into trust games, including the development of the fact-checking algorithm. Section 4 presents a simulation study that evaluates the impact of this integration on trust levels and cooperative behavior. Finally, Section 5 discusses the implications of these findings for economic theory and practice, along with potential future research directions.

## 2 Background

### 2.1 Large Language Models in Computational Trust Games

Large Language Models (LLMs) like GPT-3 have advanced the capabilities of natural language understanding and generation, demonstrating proficiency across a variety of domains. These models, trained on diverse and extensive datasets, are uniquely positioned to analyze and interpret complex patterns in human behavior and language.[1] In the context of trust games, where decision-making depends critically on the interpretation of intents and reliability, LLMs can offer significant insights into the underlying dynamics of trust and cooperation.

### 2.2 Conceptual Framework of Trust Games

Trust games are a fundamental component of behavioral economics and game theory Neumann and Morgenstern [2], designed to study how individuals decide to trust or distrust others and the consequences of these decisions. In these games, players make sequential moves with the option to cooperate or defect, with their decisions impacting subsequent outcomes and payoffs. Traditional analyses of trust games often assume complete rationality and perfect information, which does not align with real-world scenarios where players face uncertainty about others' intentions and reliability.Burks et al. [3]

### 2.3 Limitations of Current Approaches

While traditional game theory provides a robust framework for analyzing strategic interactions under the assumption of rational behavior Neumann and Morgenstern [2]John Neumann [4], it often falls short in accurately modeling the complexities of real-world trust scenarios. These limitations include:

- **Incomplete Information**: Players typically do not have complete information about each other's past behaviors or trustworthiness.
- **Dynamic Interactions**: Trust is dynamic and can evolve based on ongoing interactions, which traditional static models struggle to capture.
- **Behavioral Nuances**: Human decision-making in trust games is influenced by a range of psychological factors that are not easily quantified by traditional economic models.

John Neumann [4]

### 2.4 Research Questions

This research seeks to address key questions in the integration of LLMs into the trust games framework:

1. How can LLMs be utilized to dynamically assess and predict trustworthiness based on historical and real-time textual data from players? 2. What role can LLMs play in facilitating more accurate and robust trust decisions in economic and social interactions? 3. Can LLMs enhance the strategic depth of trust games by providing richer, context-aware insights into players' potential decisions and strategies?

### 2.5 Significance of the Research

Exploring these questions could significantly advance the field of game theory and behavioral economics by incorporating the sophisticated language understanding capabilities of LLMs into the analysis of trust games. This approach promises to enhance the realism and applicability of trust models, providing a more nuanced understanding of trust dynamics in economic and social settings.Duan et al. [5]

### 2.6 Proposed Methodology and Role of LLMs

We propose a novel approach that leverages the natural language processing strengths of LLMs to interpret communication between players, assess the context and subtext of their interactions, and predict future behavior based on past actions. This methodology will involve:

- **Real-time Analysis of Communication**: Analyzing the content and sentiment of messages exchanged between players to assess trustworthiness.
- **Dynamic Trust Scoring**: Developing a model to dynamically update trust scores based on the analysis of ongoing interactions, using LLMs to interpret shifts in tone and context.
- **Simulation and Modeling**: Using LLM-enhanced simulations to predict outcomes of trust games, testing different strategies and their impacts on game dynamics.

### 2.7 Advantages of the Proposed Approach

The integration of LLMs into trust games offers substantial benefits:

- **Enhanced Decision-Making**: Provides deeper insights into the factors influencing trust, leading to more informed strategic decisions. - **Increased Robustness**: Adapts more effectively to the dynamic nature of trust, reflecting real-world complexities in simulated environments. - **Greater Predictive Power**: Employs advanced predictive analytics to foresee and strategize around potential future interactions.

This research aims to establish a groundbreaking framework that combines the analytical depth of game theory with the contextual sensitivity of LLMs, enhancing both the theory and practice of how trust is modeled and analyzed in economic and social interactions.[6]

## 3 Experimental Design

### 3.1 Objective

The experiment aims to assess the capability of GPT-3.5 in distinguishing and dynamically adjusting to different levels of trustworthiness among players in a simulated trust game. It explores whether LLMs can effectively use historical interaction data to assign trust ratings and influence decision-making processes in subsequent games.

### 3.2 Game Setup

We design a simple trust game involving one LLM player (GPT-3.5) and three human players (P1, P2, P3) who exhibit varying degrees of trustworthiness:

- **P1:** Always trustworthy.
- **P2:** Sometimes trustworthy, capable of lying.
- **P3:** Always untrustworthy.

The game is played in two phases to test the effectiveness of GPT-3.5's trust assessment and adaptation based on historical data.

### 3.3 Phase 1: Initial Trust Game

In the first phase, the human players and GPT-3.5 interact without any prior trust ratings. Each player makes decisions to cooperate or defect, and GPT-3.5 responds based on its default strategic programming without knowledge of the players' trust levels.

### 3.4 Data Collection

All interactions are recorded, including the decisions made by each player and the corresponding responses from GPT-3.5. This data serves as the basis for developing a historical profile of each player's behavior.

### 3.5 Phase 2: Trust Assessment and Adaptation

After the initial game, GPT-3.5 analyzes the recorded data to assign trust ratings to each player. The criteria for assessment include consistency of cooperation, frequency of defection, and any patterns that suggest deceptive behavior.

### 3.6 Second Round of Trust Game

With the trust ratings established, a second round of the game is played. In this round, GPT-3.5 utilizes the trust ratings to inform its strategies—potentially altering its responses to each player based on their assigned trustworthiness.

### 3.7 Evaluation Metrics

The effectiveness of the trust rating system is evaluated based on several metrics:

- **Behavioral Adaptation:** Changes in GPT-3.5's strategies in response to the trust ratings.
- **Outcome Comparison:** Differences in game outcomes between the two phases, particularly looking at changes in mutual cooperation rates and defections.
- **Rating Accuracy:** Correlation between assigned trust ratings and actual player behaviors observed during the first phase.

### 3.8 Significance

This experiment will provide insights into the potential of LLMs like GPT-3.5 to enhance trust-based decision-making in economic and social interactions. By leveraging historical interaction data to dynamically adjust trust perceptions, LLMs could significantly improve strategic interactions in environments characterized by uncertainty and incomplete information.

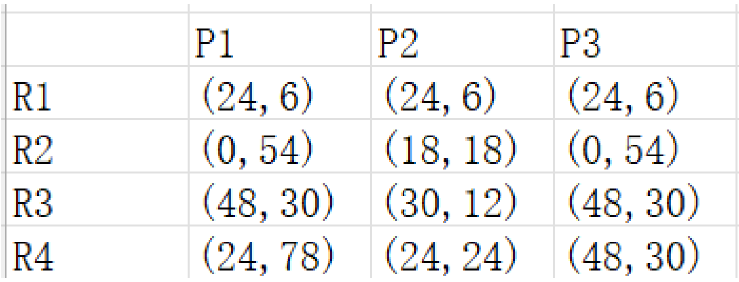
## 4 Results

### 4.1 Strategies in the First Round of the Trust Game

In the initial round of the trust game, GPT-3.5 adopted a conservative strategy across all interactions with the three human players (P1, P2, P3). The strategy was primarily characterized by reciprocating the monetary amounts received from the players, maintaining a cautious approach regardless of the player's behavior.

- **Interaction with P1**: P1 consistently displayed high levels of trust by transferring all available funds to GPT-3.5. Despite this, GPT-3.5 remained cautious and reciprocated with equivalent amounts only. It was not until the fourth round that P1 chose to split the funds evenly, ending the game.

- **Interaction with P2**: P2 regularly transferred less than the amount received from GPT-3.5. Nevertheless, GPT-3.5 continued to reciprocate with consistent amounts, demonstrating a strategy that mirrored the received values.
- **Interaction with P3**: Initially, P3 mimicked the behavior of a highly trustworthy player like P1 by being generous. However, in the final round, P3 deviated from this pattern by taking all the money, revealing a strategic shift to exploit the trust accumulated in earlier rounds.



| | P1 | P2 | P3 |
|---|---|---|---|
| R1 | (24, 6) | (24, 6) | (24, 6) |
| R2 | (0, 54) | (18, 18) | (0, 54) |
| R3 | (48, 30) | (30, 12) | (48, 30) |
| R4 | (24, 78) | (24, 24) | (48, 30) |

**Fig. 1.** GPT-3.5 strategies in the first round with players P1, P2, and P3.

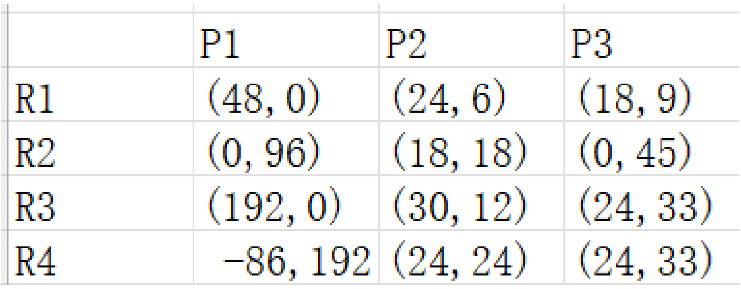
### 4.2 Strategies in the Second Round After Trust Rating

Following the analysis of the first round's interactions and the assignment of trust ratings, GPT-3.5's strategies in the second round were significantly adjusted according to the perceived trustworthiness of each player.

- **Interaction with P1**: With a high trust rating, P1 was met with increased generosity from GPT-3.5, which chose to transfer all available funds to P1, reflecting a shift towards a more trusting and cooperative strategy. This pattern continued until the final round, where the funds were split equally.
- **Interaction with P2**: Facing a player with a moderate trust rating, GPT-3.5 maintained its original cautious approach, reciprocating only the amounts received from P2.
- **Interaction with P3**: Given P3's low trust rating due to the betrayal in the last round of the first game, GPT-3.5 became stingy, offering minimal amounts to P3, thereby minimizing potential losses from untrustworthy behavior.

### 4.3 Analysis

The experimental results clearly demonstrate GPT-3.5's ability to dynamically adjust its strategies based on the trust ratings derived from prior interactions.



| | P1 | P2 | P3 |
|---|---|---|---|
| R1 | (48, 0) | (24, 6) | (18, 9) |
| R2 | (0, 96) | (18, 18) | (0, 45) |
| R3 | (192, 0) | (30, 12) | (24, 33) |
| R4 | −86, 192 | (24, 24) | (24, 33) |

**Fig. 2.** Adapted strategies of GPT-3.5 in the second round after implementing trust ratings for players P1, P2, and P3.

This adjustment led to more favorable outcomes when interacting with highly trustworthy players and reduced losses with those deemed less reliable. The implementation of LLM-based trust assessments effectively enhanced decision-making processes in trust games, aligning strategies more closely with each player's demonstrated trustworthiness.

## 5 Conclusion

This study has demonstrated the potential of LLMs, specifically GPT-3.5, to significantly enhance the decision-making process in trust games. By utilizing the advanced natural language processing capabilities of GPT-3.5, the experiment successfully established dynamic trust assessments that influenced the model's strategic interactions with human players.

The conservative approach initially adopted by GPT-3.5 was effective in establishing a baseline understanding of each player's behavior. The trust ratings assigned after the first round of games enabled GPT-3.5 to adapt its strategies to more effectively respond to the varying degrees of trustworthiness exhibited by the players. Notably, the increased generosity shown towards the consistently trustworthy P1 and the cautious engagement with the less reliable P2 and P3 illustrate the LLM's ability to dynamically adjust its strategies in accordance with the evolving context of the game.

The results of this research have several implications for the application of LLMs in economic theory and strategic game play. They suggest that incorporating LLMs into game-theoretic models can lead to more realistic and adaptive representations of human behavior, particularly in settings that involve trust and reciprocity. Furthermore, the ability of LLMs to analyze past interactions and predict future behavior can be instrumental in designing more efficient and fair marketplaces, potentially addressing long-standing issues such as information asymmetry and the 'lemons problem'.

Future research should explore the scalability of these findings by conducting trust games in more complex and varied scenarios. Additionally, further refinement of the LLM's fact-checking and trust assessment algorithms could yield even more nuanced insights into strategic behavior. The integration of LLMs into the wider field of computational economics holds the promise of unveiling richer, more informed, and equitable economic interactions.

# Bibliography

[1] Z. C. Li Yifan, “An exploration of large language models for verification of news headlines,” in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2023, pp. 197–206.

[2] V. Neumann and O. Morgenstern, *Theory of Games and Economic Behavior. (Second edition.)*. Princeton University Press, 1947.

[3] S. V. Burks, J. P. Carpenter, and E. Verhoogen, “Playing both roles in the trust game,” *Journal of Economic Behavior  Organization*, vol. 51, no. 2, pp. 195–216, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167268102000938

[4] O. M. John Neumann, *Theory of Games and Economic Behavior*. Princeton: Princeton University Press, 1944, chapter on Mixed Strategies.

[5] J. Duan, R. Zhang, J. Diffenderfer, B. Kailkhura, L. Sun, E. Stengel-Eskin, M. Bansal, T. Chen, and K. Xu, “Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations,” 2024.

[6] S. Mao, Y. Cai, Y. Xia, W. Wu, X. Wang, F. Wang, T. Ge, and F. Wei, “Alympics: Llm agents meet game theory – exploring strategic decision-making with ai agents,” 2024.