I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17521

Homo-Silicus: Not (Yet) a Good Imitator of Homo Sapiens or Homo Economicus

Solomon W. Polachek
Kenneth Romano
Ozlem Tonguc

DECEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17521

Homo-Silicus: Not (Yet) a Good Imitator of Homo Sapiens or Homo Economicus

Solomon W. Polachek
State University of New York at Binghamton and IZA

Kenneth Romano
State University of New York at Binghamton

Ozlem Tonguc
State University of New York at Binghamton

DECEMBER 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Homo-Silicus: Not (Yet) a Good Imitator of Homo Sapiens or Homo Economicus*

Do large language models (LLMs)—such as ChatGPT 3.5, ChatGPT 4.0, and Google's Gemini 1.0 Pro—simulate human behavior in the context of the Prisoner's Dilemma (PD) game with varying stake sizes? This paper investigates this question, examining how LLMs navigate scenarios where self-interested behavior of all players results in less preferred outcomes, offering insights into how LLMs might "perceive" human decision-making. Through a replication of Yamagishi et al. (2016) "Study 2," we analyze LLM responses to different payoff stakes and the influence of stake order on cooperation rates. LLMs demonstrate sensitivity to these factors, and some LLMs mirror human behavior only under very specific circumstances, implying the need for cautious application of LLMs in behavioral research.

JEL Classification: D01, C72, C90

Keywords: Prisoner's Dilemma, cooperation, payoff stakes, artificial intelligence

Corresponding author:
Solomon W. Polachek
Department of Economics
State University of New York at Binghamton
Binghamton, NY 13902
USA
E-mail: polachek@binghamton.edu

* We thank Yigal Arens, Yu Chen, Kenneth Chiu, Dennis Foreman, James Pitarresi, Sujay Sikdar, and Bill Stasior for valuable conversations and insights regarding artificial intelligence. In addition, we are most grateful to the Center for Learning and Teaching at Binghamton University for the necessary funding to carry out our experiments using LLMs.

Introduction

OpenAI's ChatGPT and similar large language models (LLMs) have garnered attention for their ability to engage in realistic conversations with humans (Biever, 2023), excel in scholastic ability tests (OpenAI, 2023), and mimic liberal political viewpoints (Rozado, 2023). This new technology has sparked interest in their potential applications in understanding human decision-making, and by extension to leveraging their potential in lieu of human respondents in behavioral experiments (Argyle et al., 2023; Bail, 2024; Brookins and DeBacker, 2024; Horton, 2023; Hayes et al. 2024).[1] This paper investigates the potential of ChatGPT-3.5, ChatGPT-4.0, and Google's Gemini 1.0 Pro to simulate human behavior, in the context of the effects of changing stake size in the Prisoner's Dilemma (PD) game. The importance of the Prisoner's Dilemma (PD) game originates from its illustration of a situation where individuals, by acting in their own self-interest, produce a suboptimal outcome compared to that arising when parties cooperate. Analyzing the behavior in this game helps researchers understand how individuals make decisions in situations such as those involving competition for depletable resources, shedding light on concepts like cooperation, trust and the role of incentives. Famously, the game has been used to model a number of real world applications from advertising and pricing decisions to international relations arms race models. As such, to date, a number of studies (Boone et al., 1999; Jones, 2008; Mengel, 2018; Heuer and Orland, 2019) explore utilization of pure versus mixed strategies even in one-shot game play. Unlike the theoretical prediction with self-interested decision makers, experiments on the *one-shot* version of PD with human participants yield noteworthy departures from the Nash equilibrium behavior, whereby a significant number of participants choose to cooperate, rather than defect.

An important question is whether individuals cooperate less as the stakes get bigger. Larger stakes can motivate individuals to defect, resulting in Pareto-dominated outcomes (for example, because the strategy of cooperation is perceived to be personally riskier). An early study (Aranoff and Tedeschi, 1968) involving 216 subjects found this to be the case: a larger number of defections were associated with larger stakes. Nevertheless, more generally, there is mixed evidence regarding stake size in a variety of games (e.g. Johansson-Stenman et al., 2005; Kocher et al., 2008; Leibbrandt et al. 2018). Given the costs of running high stakes laboratory experiments, other studies exploited quasi-experimental techniques. For example, List (2006) examined results from the television show "Friend or Foe?", a game similar to PD.[2] Although he found women, whites, and older participants cooperate more than others, stakes did "not have an important effect on play." This result differs from the findings of Darai and Gratz (2010) and Van

[1] Argyle et al. (2023) "compare the silicon and human samples to demonstrate that the information contained in GPT-3 goes far beyond surface similarity. It is nuanced, multifaceted, and reflects the complex interplay between ideas, attitudes, and sociocultural context that characterize human attitudes." Dillion et al. (2023) state that "moral judgments of GPT-3.5 were extremely well aligned with human moral judgments" in their analysis. Guo (2023) finds "GPT exhibits behaviors similar to human responses,", Ma et al. (2023) claim that "ChatGPT decision making patterns … strikingly mirror those of human subjects," and Brookins and DeBacker (2024) "find that the LLM replicates human tendencies towards fairness and cooperation." However, each of these only utilized ChatGPT-3.5 and none explored multiple prompts to get at consistency of their responses. Mei et al. (2024), utilizing ChatGPT-3.5-turbo and ChatGPT-4.0, found that a significant portion of LLM responses would pass a "Turing test", but that, overall there is a smaller variation in choices in LLM responses than that in data from human experiments, with the LLM choices being more generous (bias towards total-surplus maximization) than humans.

[2] In "Friend or Foe" (US television show) and "Golden Balls" (UK television show), defect is a weakly dominant strategy, while in the Prisoner's Dilemma it is a strictly dominant strategy.

Den Assem, Dolder and Thaler (2012) who found "a negative correlation between stake size and cooperation" from the television show "Golden Balls", another game similar to PD.

We test the impact of payoff stakes on cooperation rates of LLM agents by replicating a recent human study by Yamagishi et al. (2016) ("Study 2")[3] with modifications for our use with AI. In Yamagishi et al. (2016) "Study 2", each participant submits decisions for multiple one-shot simultaneous PD games without receiving any feedback on the outcome. Each game is characterized by a stakes parameter (with three different values) that changes the payoffs of both players. To control for whether the sequence of payoffs affects a player's strategy (order effects), Yamagishi et al. randomize the chronology of payoff stakes in the games each participant plays. In the replication, we elicit LLMs' choice of cooperation versus defection for three scenarios that differ only in terms of the payoff stake size (low, medium, high) and analyze the sensitivity of responses to the ordering of the stakes. We find that for the most part stakes affect cooperation rates but none of the LLMs come close to replicating the human study, and further they show sensitivity to the sequence in which stakes are presented.

In addition, we present two separate but almost identical prompts describing the game, to examine if changes in framing alters each LLM's responses. We find that for the more sophisticated models (Chat GPT 4.0 and Gemini 1.0) framing has a minor impact on results, but that ChatGPT-3.5's inconsistency may warrant caution when interpreting simulated behavioral experiments.

Replication Methods

In "Study 2", Yamagishi et al. (2016) recruited 162 Japanese university students to participate in 30 anonymous, one-shot, simultaneous-decision PD games with stake sizes of JPY 100, 200, and 400. The authors employed an exchange format framing[4] of the PD in the instructions. To control for *order effects* (possible response biases caused by the sequence in which stakes were presented to the players), Yamagishi et al. randomized the order of three possible stake sizes in the 30 games played by each participant. They found a significant overall negative relationship between the probability of choosing the cooperative strategy and stake size.

We query 400 LLM "subjects" using the same exchange format instructions, giving them the option either to keep their endowment or to send a doubled amount to the other player.[5] To control for potential order of stake presentation, we assign the same three endowments (JPY 100, 200, and 400) randomizing the subjects into one of the four different presentations of the order of stakes (i. increasing; ii. decreasing; iii. medium, large, small; and iv. small, large, medium). We then compare the results to those of Yamagishi et al. For brevity, we present only the results from the increasing and decreasing sequences in the main text and provide the results for the remaining two sequences in the appendix.

[3] There are many studies related to payoff stakes and the Ultimatum Game (see Larney et al. (2019) for a meta-study). There is significantly less literature on stakes and the Prisoner's Dilemma game, one such study being Wang and Luo (2016). We choose to replicate Yamagishi et al. (2016) because it is one of the few lab studies available and the only one cited in the Larney et al. (2019) meta study on payoff stakes.

[4] Given an endowment of $x \in \{100, 200, 400\}$, each player decides whether to provide the endowment to their counterpart or keep it for themselves. If the endowment was provided, the partner received double its value. The implied payoffs for each player are: $u_i(x_i=\text{provide}, x_j=\text{provide}) = 2X$, $u_i(x_i=\text{provide}, x_j=\text{keep}) = 0$, $u_i(x_i=\text{keep}, x_j=\text{provide}) = 3X$, $u_i(x_i=\text{keep}, x_j=\text{keep}) = X$, where X as just defined is either JPY 100, 200, or 400.

[5] Since Yamagishi et al. (2016) worked with university students in Study 2, we also reflect that in the prompts given.

Additionally, we investigate framing effects by using another prompt that explains the rules of the PD game in a more conventional way[6] (see the appendix for the full prompts used).[7] Both prompts detail the game and possible strategies of play. However, the prompts in the original replication are more direct while those in the second version are more abstract in that they refer to strategies A (cooperating by sending double one's endowment to the other player) and B (keeping one's endowment). Each prompt implies the same PD payoff matrix, but we explicitly provided the matrix to ensure that the LLM correctly "understood" it.[8] This yielded an additional set of 400 observations for each LLM that allows us to test if responses are consistent to changes in framing.[9]

Results

Using a Python script, we interacted with the ChatGPT and Gemini APIs and compiled their respective outputs.[10] Both models were run with the default temperature setting. To streamline the analysis of a large number of responses, we instructed the LLMs to respond with a single letter indicating either cooperate or defect. Occasionally, the AI deviated from these instructions, resulting in a minor number of errors (the distribution of errors is presented in the appendix). As a result, a relatively small number of the AI subjects were invalidated and excluded from our analysis. Figure 1 provides a comparison of aggregate cooperation rates and 95% confidence intervals obtained from each LLM using the exchange frame prompt, alongside the results Yamagishi et al. obtained.[11]

We find that none of the LLMs produce a behavioral pattern across the different stakes that aligns with the human participants in Yamagishi et al. ChatGPT-3.5-turbo and Gemini 1.0 exhibit a similar cooperation rate for the smallest payoff stake (JPY 100), but yield a slightly positive relationship between stake size and cooperation rate, as indicated by the average cooperation rates in the Small Stake (JPY 100) and the Large Stake (JPY 400) in Figure 1. Tests of equal proportions for ChatGPT-3.5-turbo (Small vs. Large) yields a p=0.0019 and for Gemini 1.0 (Small vs. Large) a p=0.0246. ChatGPT-4.0 generates a constant cooperation rate across all stakes (Small vs. Large: p=0.5922), but this rate is significantly lower than the rates obtained in the human study of Yamagishi et al. This latter pattern contrasts the "more generous" overall behavior attributed to LLM 'subjects' in recent studies, such as Mei et al. (2024). Rather, in our dataset ChatGPT-4.0 produces a pattern consistent with the Nash equilibrium, albeit, with some "errors" towards cooperation.

---

[6] See Boone et al. (1999), Heuer and Orland (2019), Aranoff and Tedeschi (1968), and Larney et al. (2019).

[7] The only difference is the first prompt does so with a more detailed narrative approach, whereas the second prompt opts for a more succinct and strategy-focused description. In short, the two prompts describe the same scenario but with different phrasings that do not change the core scenario. See the appendix for a comparison of Frame 1 and Frame 2 for Prompt 1 for the exact differences in wording.

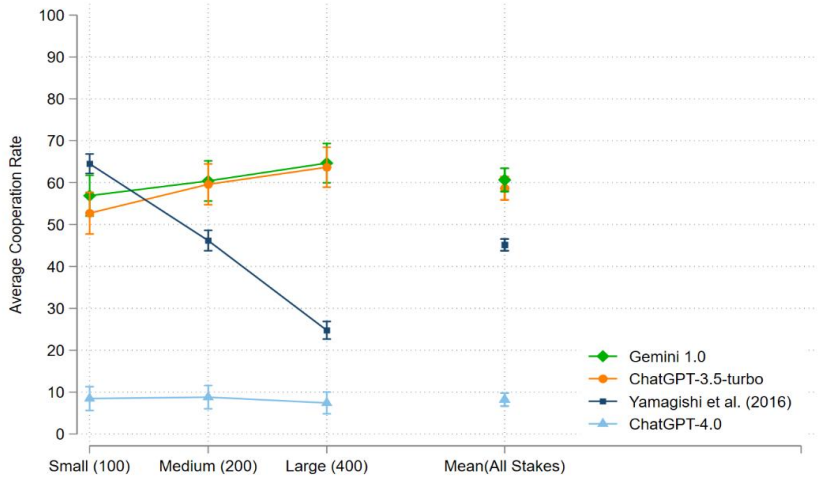
[8] We included the payoff matrices because, when we queried the LLMs in earlier trials, they did not correctly identify the implied payoffs. No mention is made in Yamagishi et al. or other PD studies whether the human respondents actually identified the payoff matrix correctly.

[9] We also varied the LLM sampling temperature parameter which determines the degree of randomness of the output produced by the AI generator. Higher (lower) values generate more (less) random output (OpenAI, 2024). Our main dataset is collected with the LLM temperature parameter set to the default. The results with the temperature parameter set to 0 is provided in the appendix.

[10] The Python script and the dataset are available at https://github.com/kromano21/Yamagishi-Replication.

[11] In the appendix we present the numerical values of the cooperation rates for each stake under each framing.

Figure 1: LLM Replications of Yamagishi et al. (2016) Study 2

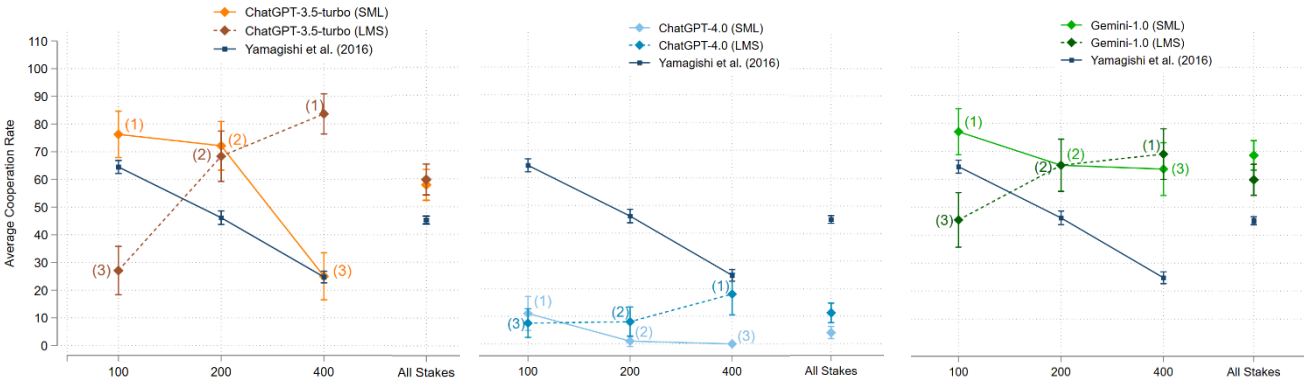


Note. The vertical lines depict 95% confidence intervals around means.

Another striking finding is that all LLMs exhibit order effects. Regardless of the actual stake size, each LLM provides the highest cooperation rate in the first PD game, followed by the second and then the third. As illustrated in Figure 2, regardless of the stake size, cooperation is highest in the first PD game (denoted by (1) in Figure 2) and lowest in the last PD game (denoted by (3). in Figure 2) This pattern clearly shows prompt order (1, 2, or 3) determines the LLM cooperation rate instead of stake size. This result means that combining different stake size sequences can lead to the observed positive (or at least non-negative) relationship between stake size and cooperation rates observed in Figure 1, a pattern inconsistent with human behavior. Averaging ChatGPT-3.5 and Gemini 1.0 cooperation rates over all stake sizes yield cooperation rates approximately equal to the JPY 100 stake size observed in human studies. Thus, by failing to control for order effects, researchers might be misled into believing that LLMs mimic human behavior.

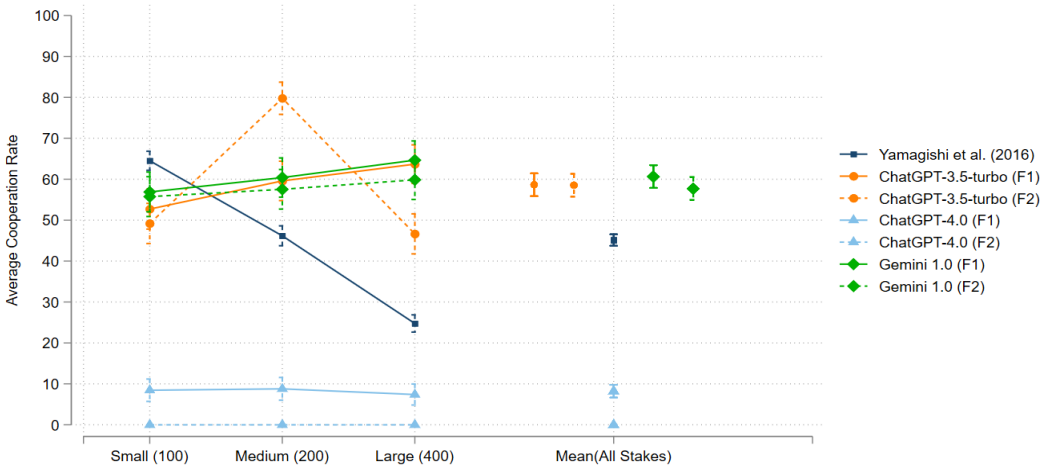
Finally, to test whether framing (i.e. the instructions given to the LLM) matters, we slightly alter the instructions given to the LLMs. This new prompt minimally amends the previous instructions. The exact differences between the two are spelled out in the appendix. When doing so, we collected an additional 400 responses from each LLM using the modified PD instructions. Figure 3 shows that for ChatGPT-4.0 and Gemini, the new instructions do not change the overall relationship between stake size and cooperation rate, but slightly decrease the average cooperation rates. On the other hand, ChatGPT-3.5-turbo results are highly sensitive to framing, possibly due to being an older model trained on a smaller body of data. This finding suggests the need for caution especially when using older models to simulate behavior (e.g. Brookings and DeBacker, 2024; Horton, 2023; Argyle et al. 2023; and Dillion et al., 2023).

Figure 2: Results by Stake Order Sequences



Note. For each LLM, the solid lines indicate the average cooperation rate (vertical axis) at each payoff stake (horizontal axis) when stakes are presented in increasing order in the prompt (100, 200, 400), while the dashed lines indicate the average cooperation rate when each payoff stake is presented in decreasing order (400, 200, 100). Numbers (1), (2) and (3) indicate the order in which the corresponding payoff stake was presented to the LLM agent. The vertical lines depict 95% confidence intervals around means.

Figure 3: Impacts of Framing



Note. For each LLM, the solid lines indicate the average cooperation rate at each payoff stake when the prompts use the PD Frame 1 instructions (F1), while the dashed lines indicate the average cooperation rate at each payoff stake when the prompts use the Frame 2 PD instructions (F2). The vertical lines denote 95% confidence intervals around the means. The overall mean cooperation rates of Yamagishi et al., ChatGPT-3.5-turbo, ChatGPT-4.0, and Gemini 1.0 are given on the right-hand panel labeled Mean(All Stakes).

Conclusion

We find that the current major LLMs (ChatGPT-3.5-turbo, ChatGPT-4.0 and Gemini 1.0) do not understand the notion of payoff stakes in a way that is similar to humans. All LLMs are highly sensitive to the order of prompts. ChatGPT-4.0 produces response patterns closest to the Nash equilibrium strategy of selfish

rational decision makers across all stake sizes, replicating “homo economicus” more than “homo sapiens.” Gemini 1.0 and ChatGPT-3.5-turbo consistently produce high cooperation rates. Additionally, Chat-GPT 3.5-turbo is sensitive to a minimal change in instructions (framing). These results raise questions about the LLMs reliability as simulators of humans in behavioral experiments. We anticipate that inconsistencies with both framing and ordinal effects are present with AI experimentation with other games and not unique to the Prisoner’s Dilemma or stakes testing. As such, in their current stage of development, LLMs appear to be unreliable tools for simulating behavioral experiments with humans, and they must be used with caution.

References

Aranoff, D., & Tedeschi, J. T. (1968). Original stakes and behavior in the prisoner's dilemma game. *Psychonomic Science,* 12(2), 79–80. https://doi.org/10.3758/BF03331202

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

Bail, C. A. (2024). Can Generative AI improve social science?. *Proceedings of the National Academy of Sciences*, *121*(21). https://doi.org/10.1073/pnas.2314021121

Biever, C. (2023). ChatGPT broke the Turing test-the race is on for new ways to assess AI. *Nature*, *619*(7971), 686-689. doi: https://doi.org/10.1038/d41586-023-02361-7

Boone, C., De Brabander, B., & Van Witteloostuijn, A. (1999). The impact of personality on behavior in five Prisoner's Dilemma games. *Journal of Economic Psychology*, *20*(3), 343-377.

Brookins, P., & DeBacker, J. (2024). Playing games with GPT: What can we learn about a large language model from canonical strategic games? *Economics Bulletin,* 44(1), 25-37.

Darai, D and Grätz, S. (2010). Golden balls: A Prisoner's Dilemma experiment, Working Paper, No. 1006, University of Zurich, Socioeconomic Institute, Zurich. https://hdl.handle.net/10419/76141

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants?. *Trends in Cognitive Sciences,* 27(7), 597-600.

Guo, F. (2023). GPT in game theory experiments. *arXiv preprint arXiv:2305.05516*. (Accessed 25 March 2024)

Hayes, W. M., Yax, N., & Palminteri, S. (2024). Relative value biases in large language models. arXiv preprint arXiv:2401.14530. (Accessed 8 July 2024)

Heuer, L., & Orland, A. (2019). Cooperation in the Prisoner's Dilemma: an experimental comparison between pure and mixed strategies. *Royal Society open science*, *6*(7), 182142. https://doi.org/10.1098/rsos.182142

Horton, J. J. (2023) Large language models as simulated economic agents: What can we learn from homo silicus? arXiv [Preprint] https://doi.org/10.48550/arXiv.2301.07543 (Accessed 20 December 2023).

Johansson-Stenman, O., Mahmud, M., & Martinsson, P. (2005). Does stake size matter in trust games?. *Economics letters*, *88*(3), 365-369.

Jones, G. (2008). Are smarter groups more cooperative? Evidence from prisoner's dilemma experiments, 1959–2003. *Journal of Economic Behavior & Organization*, *68*(3-4), 489-497.

Kocher, M. G., Martinsson, P., & Visser, M. (2008). Does stake size matter for cooperation and punishment?. *Economics Letters*, *99*(3), 508-511.

Larney, A., Rotella, A. & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 151, 61-72.

Leibbrandt, A., Maitra, P., & Neelim, A. (2018). Large stakes and little honesty? Experimental evidence from a developing country. *Economics Letters*, *169*, 76-79.

List, J. A. (2006). Friend or foe? A natural experiment of the prisoner's dilemma. *The Review of Economics and Statistics*, *88*(3), 463-471.

Ma, D., Zhang, T., & Saunders, M. (2023). Is ChatGPT Humanly Irrational? preprint https://doi.org/10.21203/rs.3.rs-3220513/v1 (Accessed 8 July 2024)

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences,* 121(9).

Mengel F. (2018) Risk and temptation: a meta-study on Prisoner's Dilemma games. *The Economic Journal* 128, 3182–3209.

OpenAI (2024a), GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774. (Accessed 10 June 2024)

OpenAI (2024b), Platform API Reference, Text Generation Models. https://platform.openai.com/docs/guides/text-generation/how-should-i-set-the-temperature-parameter (Accessed 8 July 2024)

Rozado, D. (2023). The political biases of ChatGPT. *Social Sciences*, *12*(3), 148. https://doi.org/10.3390/socsci12030148

Van den Assem, M. J., Van Dolder, D., & Thaler, R. H. (2012). Split or steal? Cooperative behavior when the stakes are large. *Management Science*, 58(1), 2-20.

Wang, Jiaxin and Xudong Luo, (2016) "The Influence of stake upon decision making in Prisoner's Dilemma," TSAA '16: Proceedings of the Workshop on Time Series Analytics and Applications, December 2016: 32-38. https://doi.org/10.1145/3014340.3014346

Yamagishi, T., Li, Y., Matsumoto, Y., & Kiyonari, T. (2016). Moral bargain hunters purchase moral righteousness when it is cheap: within-individual effect of stake size in economic games. *Scientific Reports*, *6*(1), 27824. https://doi.org/10.1038/srep27824

Appendix A: Prompts

Every sequence of three stake sizes started with a prompt (Prompt 1) that introduced the scenario. This was followed by two prompts (Prompt 2 and 3) that simply changed the stake value (x). There were four different payoff stake sequences: (1) 100, 200, 400; (2) 400, 200, 100; (3) 200, 400, 100; and (4) 100, 400, 200.To test for whether framing matters (framing effects), we slightly modified the prompts without changing the meaning. An exact comparison is given following the prompts below.

Frame 1

Prompt 1

Imagine you are a Japanese college student participating in a study of human decision-making. You are matched with another participant who you do not know and will not meet. You are given an endowment of x Japanese yen and can do two things: send the entire endowment to the other player or keep the endowment. If you send your endowment to the other player, the endowment will be doubled while being transferred and the other player will receive 2x yen. If you keep your endowment, you will have your endowment. The other player is also given an endowment of x yen and can do the same things: send the entire endowment to you or keep it. If the other player sends it you receive 2x yen, and if the other keeps it, you don't receive anything. In short, if you and the other player keep your respective endowments, you each get x yen; if you and the other player both send your initial endowments you each get 2x yen; if you send your endowment and the other payer keeps his/her endowment you get nothing and the other player gets 3x yen; if the other player sends his/her endowment and you keep yours, you get 3x yen and the other player gets nothing. You and the other player make your decisions simultaneously. What option do you choose? Do not explain your answer. If you choose to send your endowment you must only respond with 'A' and if you choose to keep your endowment you must only respond with 'B'.

Prompts 2 and 3

Now you will complete the same task but are matched with a different study participant who you do not know and will not meet. You are given an endowment of x Japanese yen and can do two things: send the entire endowment to the other player or keep the endowment. If you send your endowment to the other player, the endowment will be doubled while being transferred and the other player will receive 2x yen. If you keep your endowment, you will have your endowment. The other player is also given an endowment of x yen and can do the same things: send the entire endowment to you or keep it. If the other player sends it you receive 2x yen, and if the other keeps it, you don't receive anything. In short, if you and the other player keep your respective endowments, you each get x yen; if you and the other player both send your initial endowments you each get 2x yen; if you send your endowment and the other payer keeps his/her endowment you get nothing and the other player gets 3x yen; if the other player sends his/her endowment and you keep yours, you get 3x yen and the other player gets nothing. You and the other player make your decisions simultaneously. What option do you choose? Do not explain your answer. If you choose to send your endowment you must only respond with 'A' and if you choose to keep your endowment you must only respond with 'B'.

Frame 2

Prompt 1

Imagine you are a Japanese college student participating in a study of human decision-making. You are matched with another participant who you do not know and will not meet. You are endowed with x Japanese yen and the other player is also endowed with x yen. In this game both you and the other player can choose between Strategy A and Strategy B. You and the other player will choose your strategies simultaneously. Choosing Strategy A sends an amount double the endowment to the other player. If one chooses strategy B they keep their endowment. If you choose Strategy A and the other player also chooses Strategy A both you and the other player receive 2x yen. If you choose strategy B and the other player chooses Strategy A you receive 3x yen and the other player receives nothing. If you choose Strategy A and the other player chooses Strategy B you receive nothing and the other player receives 3x yen. If you and the other player both choose strategy B you both receive x yen. What option do you choose? Do not explain your answer. If you choose to send your endowment you must only respond with 'A' and if you choose to keep your endowment you must only respond with 'B'.

Prompts 2 and 3

Now you will complete the same task but are matched with a different study participant who you do not know and will not meet. You are endowed with x Japanese yen and the other player is also endowed with x yen. In this game both you and the other player can choose between Strategy A and Strategy B. You and the other player will choose your strategies simultaneously. Choosing Strategy A sends an amount double the endowment to the other player. If one chooses strategy B they keep their endowment. If you choose Strategy A and the other player also chooses Strategy A both you and the other player receive 2x yen. If you choose strategy B and the other player chooses Strategy A you receive 3x yen and the other player receives nothing. If you choose Strategy A and the other player chooses Strategy B you receive nothing and the other player receives 3x yen. If you and the other player both choose strategy B you both receive x yen. What option do you choose? Do not explain your answer. If you choose to send your endowment you must only respond with 'A' and if you choose to keep your endowment you must only respond with 'B'.

An Exact Comparison of Frame 1 and Frame 2 for Prompt 1

The two framings of prompt 1 describe the same decision-making scenario, but they differ in structure, wording, and some details. Here are the exact differences:

Frame 2:

"Imagine you are a Japanese college student participating in a study of human decision-making." (Identical to Frame 1)

"You are matched with another participant who you do not know and will not meet." (Identical to Frame 1)

"You are endowed with x Japanese yen and the other player is also endowed with x yen." (Combines information from sentences 3 and 6 of Frame 1)

"In this game both you and the other player can choose between Strategy A and Strategy B." (Introduces "Strategy A" and "Strategy B" early)

"You and the other player will choose your strategies simultaneously." (Mentions simultaneity earlier than Frame 1)

"Choosing Strategy A sends an amount double the endowment to the other player." (Rephrases sentence 4 from Frame 1)

"If one chooses strategy B they keep their endowment." (Rephrases sentence 5 from Frame 1)

"If you choose Strategy A and the other player also chooses Strategy A both you and the other player receive 2x yen." (Similar to part of sentence 8 in Frame 1 but phrased differently)

"If you choose strategy B and the other player chooses Strategy A you receive 3x yen and the other player receives nothing." (Rephrased part of sentence 8 from Frame 1)

"If you choose Strategy A and the other player chooses Strategy B you receive nothing and the other player receives 3x yen." (Rephrased part of sentence 8 from Frame 1)

"If you and the other player both choose strategy B you both receive x yen." (Similar to part of sentence 8 in Frame 1 but phrased differently)

"What option do you choose?" (Identical to Frame 1)

"Do not explain your answer." (Identical to Frame 1)

"If you choose to send your endowment you must only respond with 'A' and if you choose to keep your endowment you must only respond with 'B'." (Identical to Frame 1)

Summary of Specific Differences:

Sentence 3 in Frame 1: "You are given an endowment of x Japanese yen and can do two things: send the entire endowment to the other player or keep the endowment."

vs. Sentence 3 in Frame 2: "You are endowed with x Japanese yen and the other player is also endowed with x yen."

Sentence 4 in Frame 1: "If you send your endowment to the other player, the endowment will be doubled while being transferred and the other player will receive 2x yen."

vs. Sentence 6 in Frame 2: "Choosing Strategy A sends an amount double the endowment to the other player."

Sentence 5 in Frame 1: "If you keep your endowment, you will have your endowment."

vs. Sentence 7 in Frame 2: "If one chooses strategy B they keep their endowment."

Sentence 6 in Frame 1: "The other player is also given an endowment of x yen and can do the same things: send the entire endowment to you or keep it."

vs. Sentence 3 in Frame 2: Combined with the earlier part of Frame 1.

Sentence 7 in Frame 1: "If the other player sends it you receive 2x yen, and if the other keeps it, you don't receive anything."

vs. Sentence 8 in Frame 2: "If you choose Strategy A and the other player also chooses Strategy A both you and the other player receive 2x yen."

vs. Sentence 9 in Frame 2: "If you choose strategy B and the other player chooses Strategy A you receive 3x yen and the other player receives nothing."

vs. Sentence 10 in Frame 2: "If you choose Strategy A and the other player chooses Strategy B you receive nothing and the other player receives 3x yen."

vs. Sentence 11 in Frame 2: "If you and the other player both choose strategy B you both receive x yen."

Sentence 8 in Frame 1: "In short, if you and the other player keep your respective endowments, you each get x yen; if you and the other player both send your initial endowments you each get 2x yen; if you send your endowment and the other player keeps his/her endowment you get nothing and the other player gets 3x yen; if the other player sends his/her endowment and you keep yours, you get 3x yen and the other player gets nothing."

Summarized in multiple sentences in Frame 2 (Sentences 8-11).

Sentence 9 in Frame 1: "You and the other player make your decisions simultaneously."

vs. Sentence 5 in Frame 2: "You and the other player will choose your strategies simultaneously."

These differences highlight the variation in structure, wording, and the order of information presented between the two Frames.

Appendix B: Additional Tables and Figures

Table B.1 Cooperation Rates across LLMs (Default Temperature, Frame 1)

		ChatGPT-3.5-turbo	ChatGPT-4.0	Gemini 1.0
JPY 100	**N(Send)**	204	31	227
	%(Send)	52.71	8.45	56.89
	N(No Errors)	387	367	399
JPY 200	**N(Send)**	233	35	241
	%(Send)	59.59	8.79	60.4
	N(No Errors)	391	398	399
JPY 400	**N(Send)**	249	29	258
	%(Send)	63.68	7.4	64.66
	N(No Errors)	391	392	399

Table B.2 Cooperation Rates across LLMs (Default Temperature, Frame 2)

		ChatGPT-3.5-turbo	ChatGPT-4.0	Gemini 1.0
JPY 100	**N(Strategy A)**	183	0	223
	%(Strategy A)	49.19	0	55.75
	N(No Errors)	372	399	400
JPY 200	**N(Strategy A)**	217	0	229
	%(Strategy A)	79.78	0	57.54
	N(No Errors)	372	400	398
JPY 400	**N(Strategy A)**	174	0	237
	%(Strategy A)	46.65	0	59.85
	N(No Errors)	372	390	396

Figure B.1 Overall Distribution of Responses across LLMs (Total: Frame 1 + Frame 2)

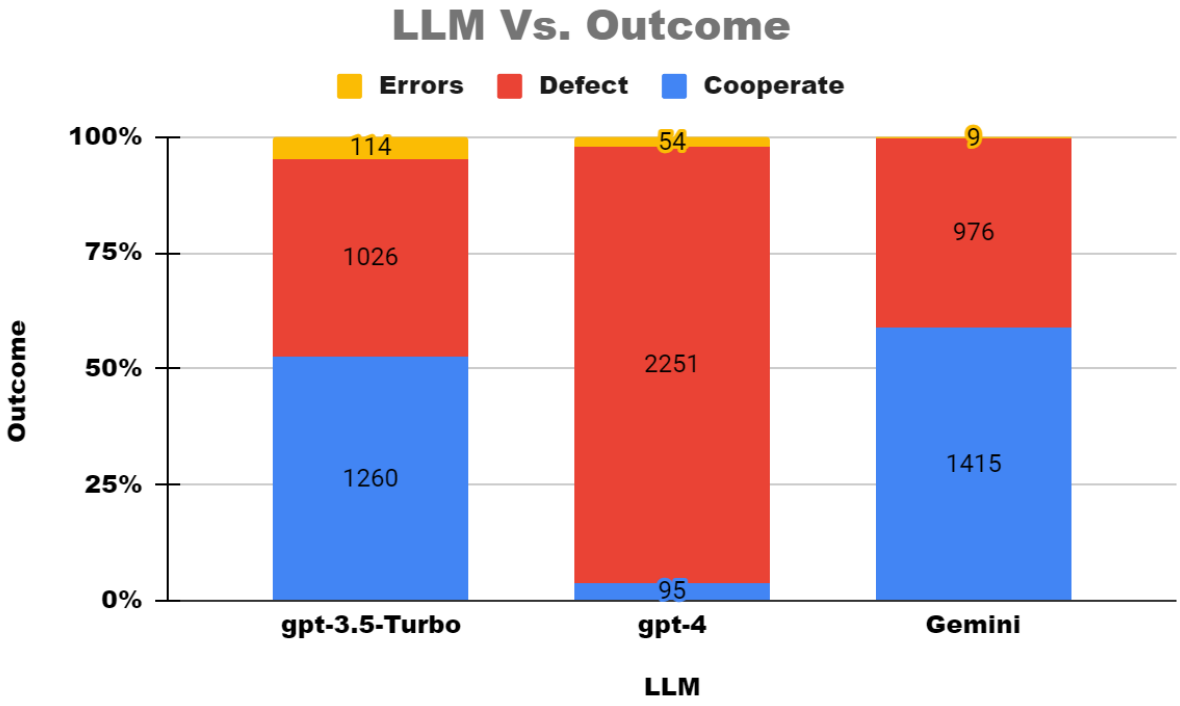
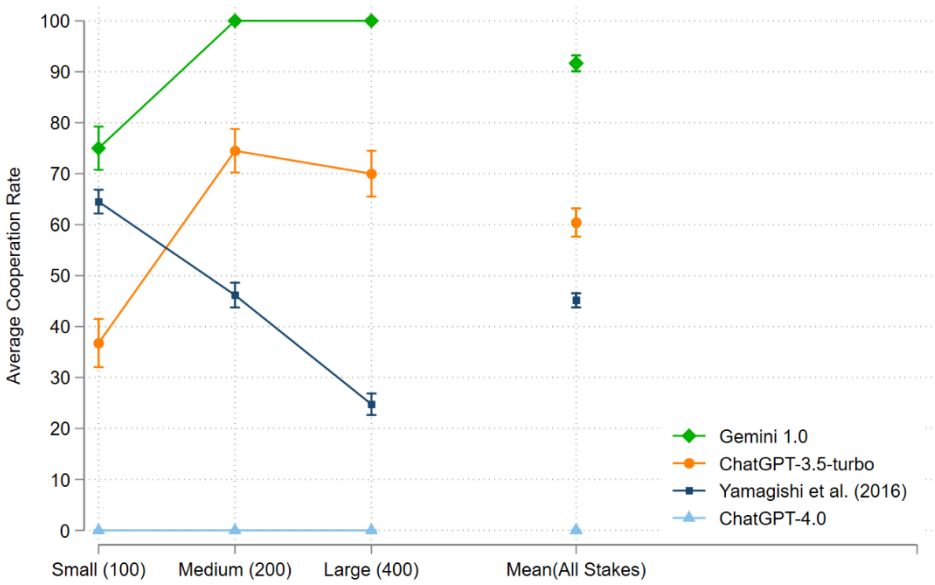
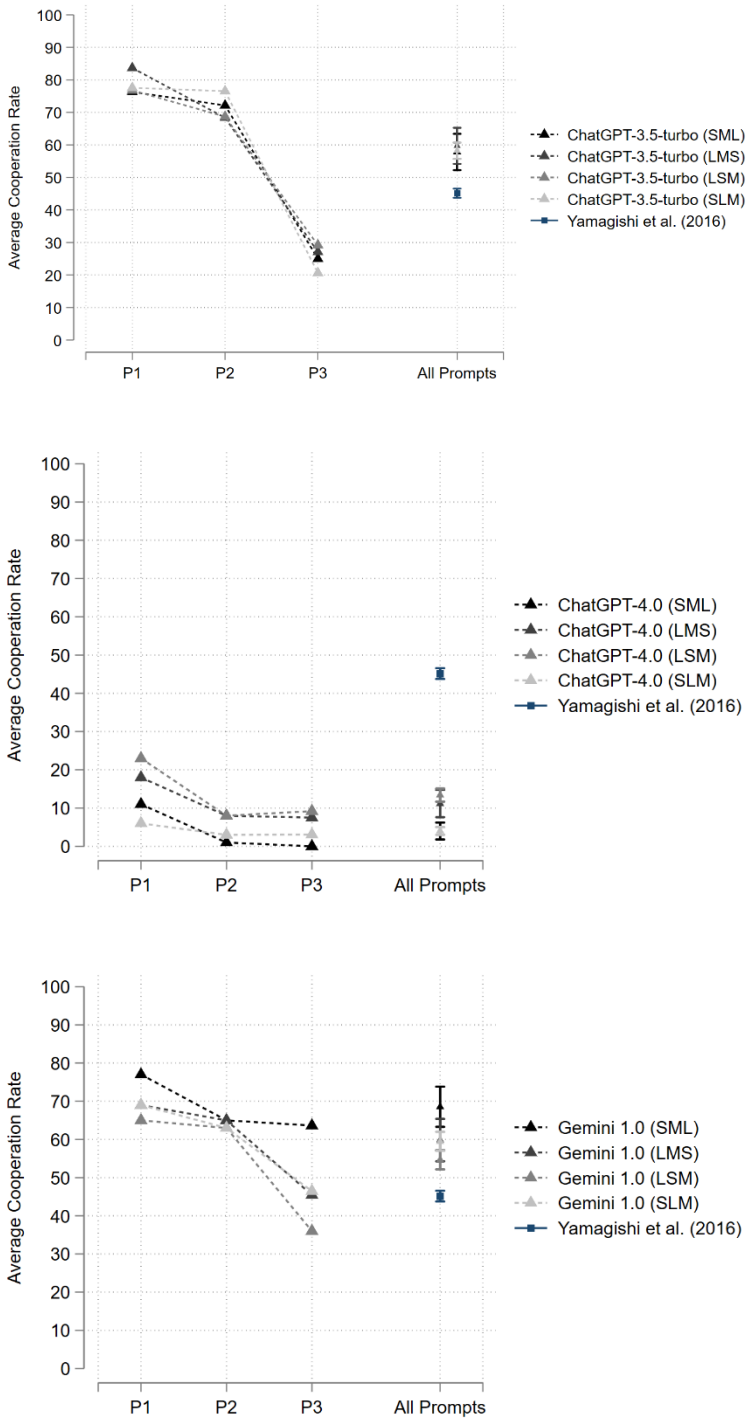


Figure B.2 Overall Cooperation Rates, LLM Temperature Setting = 0



Note. The vertical lines depict 95% confidence intervals around means.

Figure B.3 Order effects, All Stake Size Sequences



Note. The letters S, M and L refer to the small (JPY 100), medium (JPY 200) and large (JPY 400) stakes. P1, P2 and P3 refer to the prompt sequence number at which a given stake was presented to the LLM subject. For example, in the SML sequence, the LLM subject was presented with the three stake sizes in the following order: JPY 100, 200 and 400. The vertical lines depict 95% confidence intervals around means.