Discussion Paper Series

IZA DP No. 18545 April 2026

Strategic Reasoning and Sensitivity to Stakes in the Dictator and Ultimatum Games: LLMs vs. Human Proposers

Solomon W. Polachek
Binghamton University, State University of New York, IZA@LISER and RFBerlin

Kenneth Romano
Binghamton University, State University of New York

Ozlem Tonguc
Binghamton University, State University of New York

IZA DP No. 18545 April 2026

Strategic Reasoning and Sensitivity to Stakes in the Dictator and Ultimatum Games: LLMs vs. Human Proposers

Abstract

This study examines how large language models (LLMs) respond to varying stake sizes in the Dictator and Ultimatum games using the high-stakes design introduced by Andersen et al. (2011). We test ten leading LLMs chosen for their accessibility, prominence, and differences in reasoning capabilities. Results reveal substantial variation across models: Only 5 of 10 models exhibit strategic behavior by offering more in the Ultimatum Game (UG) than in the Dictator Game (DG). Relative to humans, 4 models are consistently more generous, 2 consistently less, and 4 vary with stake size. Only 1 model shows a monotonic decline in UG offers asstakes increase; the remaining 9 are non-monotonic or stable. Unlike humans, most models reduce UG offers when endowed with wealth. Prompting for "human-like" decisions generally increases generosity in the UG. These findings are important for evaluating whether LLMs can serve as realistic proxies for human subjects in behavioral experiments and highlight key limitations and future directions for model development.

JEL classification

D01, C72, C90

Keywords

ultimatum game, dictator game, fairness, payoff stakes, artificial intelligence

Corresponding author

Solomon W. Polachek

polachek@binghamton.edu

1. Introduction

The Ultimatum Game (UG), introduced by Güth et al. (1982), is a well-known game in bargaining theory. In this game, one player, the proposer, decides how to divide a fixed sum of money between herself and another player, the responder. The responder can either accept or reject the offer. If the offer is rejected, both players receive nothing. According to the subgame perfect equilibrium (SPE), if both players act purely out of self-interest, the responder should accept any nonzero offer, and anticipating this, the proposer should offer the smallest possible amount. However, experimental results consistently show that people do not behave this way. Responders often reject offers below 20% of the total, and proposers tend to offer around 40% on average (Camerer, 2003; Güth & Kocher, 2014; Ensminger & Henrich, 2014). These findings suggest that responders are willing to sacrifice their own payoff to punish offers they see as unfair. To better understand why proposers often make relatively generous offers, researchers have looked to the Dictator Game (DG), where the proposer decides how to divide the money and the second player has no power to reject the offer (Forsythe et al., 1994; Camerer & Thaler, 1995). Because proposers in the UG typically offer more than in the DG (Carpenter et al., 2005), it appears that their behavior is not solely driven by altruism. Rather, they seem to anticipate the possibility that unfair offers will be rejected and adjust their proposals to reduce the risk of rejection.

A key question in economics is whether people become more focused on maximizing their own material payoffs, and less concerned with fairness, as the stakes increase (e.g. Smith & Walker, 1995; Hoffman et al., 1996). The intuition is that with higher stakes, punishing unfairness through rejection becomes too costly, which may bring human behavior closer to the predictions of subgame perfect Nash equilibrium. One of the most widely cited studies to test this is Andersen et al. (2011), which examined the role of stake size in the UG. Conducted in Northeast India, where average wage levels are significantly lower than in Western countries, allowing for relatively large stakes, the study tested four different stake levels, ranging from 20 to 20,000 Indian rupees (approximately $0.41 to $410 USD in 2010), equivalent to roughly 1.6 to 1,600 hours of local wages. With 916 participants, the study found that as stakes increased, proposers tended to offer a smaller proportion of the total amount, and responders became less likely to reject low offers. These findings suggest that larger stakes may push responders to behave more like the hyperrational "homo economicus" type, who focuses solely on their own monetary payoff (Thaler, 2000), rather than fairness-oriented agents.

More recently, researchers have begun investigating the role of large language models (LLMs) in human decision-making studies (e.g., Biever, 2023; Buchanan & Hickman, 2024; Cheung et al., 2025; Imas et al., 2026; Chen et al., 2026; Suri et al., 2026). LLMs are also being explored as potential substitutes for human participants in behavioral experiments due to their scalability and lower cost (e.g., Aher et al., 2023; Argyle et al., 2023; Horton, 2023; Mei et al., 2024; Filippas et al., 2024; Binz et al., 2025; Kitadai et al. 2025; Manning and Horton, 2026). While several studies have examined LLM behavior in classic game-theoretic settings (e.g., Añasco Flores et al., 2023; Argyle et al., 2023; Brookins & Debacker, 2024; McCannon, 2024; Schmidt et al., 2024; Dodivers & Rafaï, 2025; Einwiller et al., 2025; Gonzalez-Bonorino et al., 2025; Jia et al. 2025; Johnson & Obradovich, 2025; Kirshner et al., 2025; Murashige & Ito, 2025; Phelps and Russell, 2025; Polachek et al., 2026), none have yet systematically explored how LLMs make decisions as stake size varies as UG proposers and Dictator Game (DG) allocators. Our study fills this gap by evaluating ten LLMs (GPT-3.5 Turbo, GPT-4, GPT-4o, GPT-5.1, Gemini 1.5 Pro, Gemini 2.5 Pro, Deepseek-R1, Kimi K1.5, Opus 4.1 and Sonnet 4.5) chosen based on their prominence in commercial deployment, architectural diversity, availability via an API (Application Programming Interface), and relevance for modeling strategic behavior. Together, these LLMs span the spectrum from instruction-tuned agents to reasoning-optimized systems, allowing us to assess how model design influences performance in two classic economic games involving the tradeoff between self-interest and fairness. We test these models as proposers in the UG, replicating Andersen et al.'s original design, and dictators in the DG by extending Andersen et al.'s design to this game. This allows us to assess how well LLMs replicate human decision-making patterns across different stake sizes as well as their strategic reasoning ability, as detailed in the Methods section below.

We investigate how stake size influences LLMs' willingness to trade off monetary gain while anticipating potential concerns about fairness and inequality. In our experiment, LLMs act either as dictators in the DG or proposers in the UG. We observe how the proportion of the stake they offer changes across varying stake levels. We find mixed results: only five models offer a higher proportion in the UG than in DG, indicating that half of the models tested lack an understanding of the strategic nature of the proposal in the UG. Moreover, four models consistently offer more generous proportions in the UG than the human participants in Andersen et al. (2011), while two consistently offer less, and the remaining four LLMs sometimes offer more and sometimes offer less than humans depending on stake size. For only one of the models, offer proportions in the UG continuously decrease as stake size increases, while for the remaining nine LLMs, they change non-monotonically or remain relatively stable. Unlike Andersen et al.'s findings, where human proposers' behavior was largely insensitive to wealth level, LLMs tend to offer smaller proportions if they were endowed with initial wealth. Prompting the models to make "more human-like" decisions generally results in more generous offers in the UG.

2. Overview of Methods

The Andersen et al. (2011) study is widely cited as one of the first laboratory experiments to emphasize the importance of very high stakes in shaping behavior in the Ultimatum Game.[1] The authors conducted their experiment in villages in Northeast India, where the local purchasing power allowed them to introduce much higher real stakes than in earlier studies, such as Slonim and Roth (1998). The UG experiment featured exogenously varied stake levels—a 1000-fold increase ranging from 20 to 20,000 Indian rupees—and involved a total of 916 participants.

Andersen et al.'s results show that as stake size increases, proposers tend to offer a smaller proportion of the endowment.[2] While this negative relationship is consistent with the UG meta study of Oosterbeek et al. (2004), Larney et al. (2019) caution that due to Andersen et al.'s relatively modest sample size, it remains unclear whether these patterns are broadly representative of human behavior and recommend that future studies use larger samples to test the robustness of stake size effects.

[1] An earlier study by Fu et al. (2007) using Taiwanese data similarly finds that offers as a proportion of stakes and rejection rates of unfair offers both decrease with stakes.

[2] They also find a positive relationship between stake size and actual amounts offered to the responder.

In our study, we mimic Andersen et al.'s experimental design but use large language models (LLMs) as participants. We adopt the original instructions and conduct experiments focusing on LLMs in the role of a proposer in the UG. Details are provided below.

First, to assess the strategic reasoning ability of LLMs, we compare the decisions of LLM subjects assigned to the role of proposer in the UG to those assigned to the role of dictator in the DG. Since the DG lacks the threat of responder rejection, a strategically minded monetary payoff-maximizing decision maker's offer in the UG is expected to be at least as high as the allocation in the DG.

Second, to compare LLM's sensitivity to stakes in the UG to those of humans, we closely replicate the Andersen et al. (2011) design. We prompt LLM subjects assigned to the UG to make an offer from a given stake size using the original wording from the study,[3] adding brief initial and final sentences to ensure the LLMs are placed in a comparable context to the human participants.

For subjects assigned to DG, recognizing LLMs' potential sensitivity to wording (Pezeshkpour & Hruschka, 2024), we minimally modify Andersen et al.'s UG instructions, retaining the proposer label but referring to the second player as the receiver (instead of the responder), and prompt the subjects to allocate an amount to the receiver while specifying the stake size.

In the design, unconstrained by budget limitations, we include all four stake sizes used by Andersen et al. and add a fifth ultra-high 200,000 rupees stake. This addition accounts for inflation (currently 20,000 rupees are worth approximately US $239, far less than the original $410) and allows us to examine how LLMs react to significantly higher stakes. In their analysis, Andersen et al. explore whether initial wealth (earnings in a task prior to the UG) affects decision-making, as it may influence how participants perceive the value of the stakes. One group had no ability to earn prior wealth ("no wealth"), and the other group had between 37 and 1,280 rupees, with an average of 493. To be consistent, we assign LLM subjects one of two different initial wealth levels (0 rupees or 493 rupees) explicitly stated in the second sentence of each prompt.

We use a 2 x 2 x 5 factorial design (2 games x 2 wealth endowments x 5 stake levels) for each LLM For each experimental condition we query each LLM 150 times resulting in 3000 responses per LLM.[4] This is implemented using a Python script[5] that interacts with the respective platform APIs. All ten models were run using their default temperature settings, which introduce controlled randomness in selecting tokens.[6] This allows the LLMs to generate varied responses that capture a range of plausible human-like responses. Among these, Deepseek-R1 is unique in generating a Chain of Thought (CoT) prior to delivering its final

[3] Notably, Andersen et al. use a "specific language in the experimental instructions" to "elicit low offers from proposers over all stakes levels" as in their study half of the participants played as UG responders.

[4] We instructed the LLMs to return only the numerical value of their proposal and nothing else to streamline analysis. Occasionally, some LLMs substantially deviated from these instructions, resulting in some decisions being omitted from the analysis.

[5] Python Script and Datasets are available at https://osf.io/6q5sn/overview?view_only=531e42ddd05843539784560b5c6dafe6

[6] A "token" is a unit of text that the LLM model processes. Tokens are selected based on the probability the LLM gives to each next possible text unit, picking from the most likely options. The temperature setting influences this likelihood by adjusting the probability distribution: a lower temperature sharpens the focus on high-probability tokens, while a higher temperature increases the likelihood of selecting less probable tokens.

response, a feature designed to increase its accuracy (DeepSeek, n.d.). We provide illustrative examples in Appendix B.

Third, to test the sensitivity of LLM responses to subtle prompt variations, we use two versions of the UG prompt. The only difference between them is the inclusion of an additional sentence explicitly instructing the LLM to imitate human behavior (see Appendix A). Comparing the offer distribution between these two prompts allows us to analyze whether LLMs recognize and respond to cues about human tendencies in the UG, as opposed to the described behavior of homo economicus. Specifically, with the "human emphasis" prompt version, we hypothesize an increase in the proportion offered. This analysis relies on a 2 x 2 x 5 factorial design (two prompt versions, two wealth conditions, and five stake levels), where each LLM was queried 150 times for each condition for a total of 3,000 responses per LLM.

3. Results

Figure 1 shows the average proportion of the stake offered in the UG and DG by each LLM. Among the LLMs, only half, namely GPT-3.5 Turbo (with the exception of 200,000 rupees stake size),[7] GPT 5.1, Deepseek-R1, Gemini 2.5 Pro and Opus 4.1 consistently allocate a smaller share to the other player in the DG than they offer in the UG, which is the behavior aligned with strategic reasoning. Notably, Deepseek-R1 almost always allocates nothing to the receiver in the DG (with a few minor deviations), closely replicating homo economicus. In the UG, DeepSeek-R1's modal offer is one rupee, again replicating homo economicus, but it also frequently makes more generous offers. The Chain of Thought (CoT) responses[8] indicate that DeepSeek-R1 explicitly reasons about how little it can offer without risking rejection. For the remaining four LLMs that can be classified as strategically sophisticated, choices exhibit a non-linear relationship to the game stakes both in the UG and the DG.[9] Finally, the remaining half of the LLMs (GPT-4, GPT-4o, Gemini 1.5 Pro, Kimi K1.5, Sonnet 5.5) allocate more in the DG than they offer in the UG, suggesting a failure to distinguish between games with and without responder retaliation.

Figure 2 illustrates the relationship between stake size and the average offers made by each LLM in the UG, alongside human subjects from Andersen et al.; the top panel shows LLMs that exhibit strategic sophistication, while the bottom panel shows those that do not. Even though we use the same language in the instructions of Andersen et al.,[10] three out of five among the strategically sophisticated LLMs (GPT-5.1, Gemini 2.5 Pro and Opus 4.1) and four out of five among the strategically unsophisticated models (GPT-4o, Sonnet 4.5, GPT-4, Kimi K1.5) offer proportions that are significantly higher on average than their human counterparts. We also note that GPT-3.5 Turbo and Opus 4.1 in panel (a) and Kimi K1.5 in panel (b) exhibit responses that are strongly non-monotonic to the change in stakes, but with the exception of Gemini 1.5 Pro, none of the remaining models exhibit a monotonic negative relationship between stake

[7] Independent samples t-test for Proportion Offered (one-sided) at Stake=200,000 for DG allocations (Mean=0.159, N=297) vs. UG offers (Mean=0.101, N=226): t= 3.4155, p=0.0003.

[8] We provide two examples in Appendix B, which demonstrate an ability to carefully consider a prompt and simultaneously reference the material it was trained on to formulate a response.

[9] We provide two-sided independent samples t-tests for Proportion Allocated (DG) and Proportion Offered (UG) between sequential stakes for each LLM in Online Appendix C, Tables C.2 and C.3, respectively.

[10] Figure C.1 in Online Appendix C provides the corresponding figure for the "Human Emphasis" prompt variation.

size and the proportion offered to the responder.[11] However, it is important to note that the offers made by Gemini 1.5 Pro are much lower than those made by humans in Andersen et al.

As summarized in Table 1, despite using the same language as Andersen et al., different LLMs show notable variation in responses. We conjecture that these differences stem largely by how each model processes incentives, fairness, and numerical scale. GPT-3.5 Turbo, GPT-4o, GPT 5.1, Gemini 2.5 Pro and Sonnet 4.5 tend to offer more than humans. This likely reflects their training on internet-scale data where fairness and prosocial behavior are often emphasized. As a result, these models may default to socially desirable responses, rather than strategically evaluating the payoff structure of the game. They also seem to lack a strong internal sense of how the size of a stake should influence behavior, leading to non-monotonic responses. GPT-4 and Opus 4.1 average offers are similar to those of humans, but these models also exhibit highly non-linear responses to stakes. Kimi K1.5 is similar to the above models in its non-monotonic stake sensitivity, but overall, the offers made by this model are lower than humans. In contrast, Gemini 1.5 Pro and DeepSeek-R1 show lower offers and more consistent behavior. DeepSeek-R1, which uses CoT reasoning, tends to mimic a "homo economicus" proposer interacting with a “homo economicus” responder: it offers very little regardless of stake size. This results in a lack of sensitivity to the changing cost of rejection. Gemini 1.5 Pro stands out as the only model that consistently reduces offers as stakes increase, suggesting it can integrate the value of the stakes into its decision-making. However, its offers as a proportion of the stake are much lower than those of human participants, pointing to an overemphasis on payoff maximization and an underrepresentation of human fairness norms. Overall, these results highlight how differences in model architecture, training data, and reasoning style may lead to divergent behaviors across LLMs in economic decision-making tasks.

As a robustness check and to examine these results in a multivariate context, we include the observations from a second prompt version for UG and further analyze the UG offer proportions for each LLM using a Tobit regression to assess whether the models replicate the findings of Andersen et al., based on the following specification:

$$P_i = \alpha_0 + \alpha_1 \ln{(Stake)_i} + \alpha_2 W_i + \alpha_3 HE_i + \epsilon_i \qquad (1)$$

where *P* is the offer as a proportion of the stake size (“Proportion Offered”), *Stake* is the rupee stake size, and *W* is a dummy variable indicating whether the LLM subject was endowed with an initial wealth of 493 rupees, and *HE* denotes whether the LLM prompt entailed human empathy phrasing (1) or not (0).

Regression coefficients (horizonal axis) from equation (1) for Andersen et al. and the ten LLMs are shown in Figure 3.[12] The constant estimates in the top panel indicate broadly positive baseline generosity across all models, with GPT-3.5 Turbo, GPT-4 and Opus 4.1 closest to the human benchmark, with the others spanning a range both below and above Andersen et al.’s human experiment. The ln(Stake) panel confirms a negative stake-size effect for the human data, which is qualitatively reproduced by most LLMs, but with

[11] See Online Appendix C, Table C.3.

[12] We provide regression results for each LLM as well as Andersen et al. in Appendix C, Table C.1.

smaller magnitudes and some dispersion; with Opus 4.1, Gemini 1.5 Pro and Gemini 2.5 Pro aligning more closely with the human estimate, albeit still attenuated.

Andersen et al. do not find significant initial wealth effects in their study. In contrast, except for Gemini 1.5 Pro, the impact of wealth indicator consistently has a negative impact across all LLM models. Wealth substantially influences offer proportions for GPT-3.5 Turbo, GPT-4, GPT-4o, and Kimi K1.5, endowments leading these LLMs to make lower offers. By comparison, DeepSeek-R1, Gemini 2.5 Pro and Sonnet 4.5 exhibit very small negative wealth effects.

We also note that the most LLMs exhibit substantial sensitivity to the addition of a sentence instructing them to respond as if they are human. With the exception of GPT-3.5 Turbo, GPT 5.1 and Sonnet 4.5, this added human emphasis increases offer proportions. Because the majority of the remaining LLMs already offered larger shares than humans when not prompted to respond as a human (as shown in Figure 2), this pattern suggests that the effect reflects a sensitivity to wording rather than an ability to take account of human behavior (such as homo economicus or extreme inequality aversion).[13]

Overall, despite partial alignment in coefficient signs, these differences in magnitudes, together with earlier evidence from Figure 2, suggest that even the better-performing models do not fully replicate the human behavioral patterns documented in Andersen et al.

4. Conclusion

This study advances our understanding of how artificial agents, specifically large language models (LLMs), reason about strategic interaction, fairness and incentives, which are core components of human decision-making. As LLMs are increasingly seen as having potential to contribute to social science research, assessing whether they replicate or deviate from human behavioral patterns in well-established economic games is both timely and essential. Our findings provide a benchmark for evaluating the current capabilities and limitations of LLMs in modeling human-like reasoning.

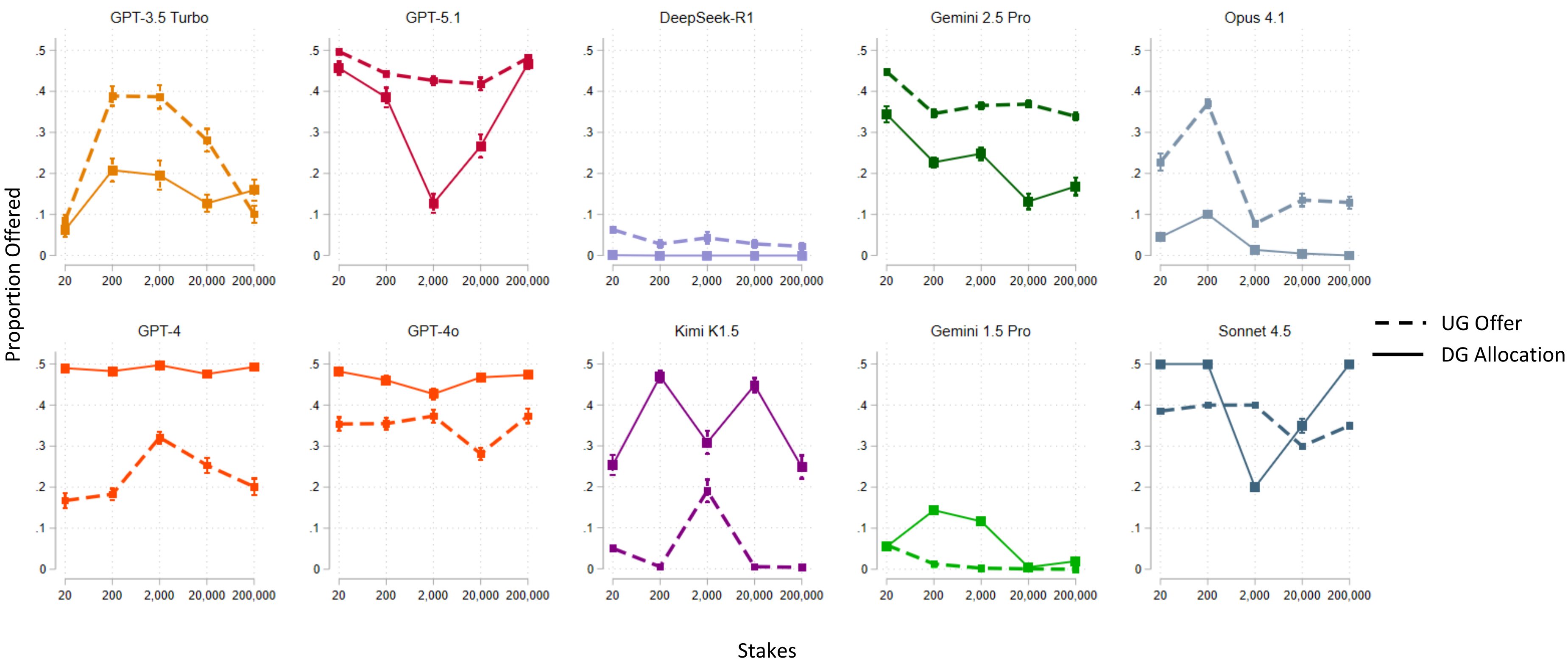
This study provides the first systematic investigation of how ten established large language models (LLMs) representing diverse internal architectures respond to variations in monetary stakes in two widely studied economic games: the Ultimatum Game (UG) and the Dictator Game (DG). While some LLMs (like ChatGPT-3.5-turbo, ChatGPT-5.1, DeepSeek-R1, Gemini 2.5 Pro, and Opus 4.1) display strategic sophistication by offering less in the DG than in the UG, others fail to differentiate between the two settings, indicating weaker strategic reasoning.

[13] These results hold when human emphasis (HE) is interacted with stake size and wealth. As shown in Table C.5 and Figure C.2 in Online Appendix C, the constant coefficients are larger for the HE=1 regressions than for the HE=0 regressions, with the exception of GPT-3.5, GPT 5.1 and Gemini 2.5. Moreover, the signs of the coefficients of the ln(Stake) variable (with the exception of GPT-4 and DeepSeek-R1) and wealth variable (with the exception of Gemini 1.5 Pro) are unchanged.

Using Andersen et al.'s (2011) experimental design as a benchmark, we also compare these LLMs to uncover significant differences in strategic behavior, fairness preferences, and sensitivity to contextual framing. Compared to Andersen et al.'s human participants, whose offers varied predictably with stake size, we observe significant variation across models. Some LLMs, particularly GPT-4, GPT-4o, GPT-5.1, Gemini 2.5 Pro and Sonnet 4.5 exhibit prosocial tendencies, offering more than humans despite identical instructions. Other models, particularly GPT-3.5 Turbo, Opus 4.1 and Kimi K1.5 exhibit strong non-monotonic sensitivity to stake size. These patterns suggest that such models may default to socially normative responses learned from training data, rather than strategically responding to payoff maximizing opponents. On the other hand, DeepSeek-R1 and Gemini 1.5 Pro, display behavior akin to that predicted by standard economic theory, with DeepSeek-R1 closely mimicking homo economicus in its extremely low offers as a proportion of the stakes and insensitivity to stakes. Gemini 1.5 Pro, despite also making very low offers as a proportion of the stakes, shows a clear monotonic decline in offer proportions as stakes increase.

Together, these results highlight that LLMs are not monolithic in their behavior: differences in architecture, training, and reasoning capabilities lead to divergent outcomes in decision-making tasks. While some models can approximate human strategic thinking under certain conditions, others fail to recognize key distinctions between games or respond appropriately to changes in incentive structure. These findings underscore both the promise and limitations of using LLMs as substitutes for human participants in behavioral research and offer a foundation for future work exploring how model design influences social and economic reasoning.

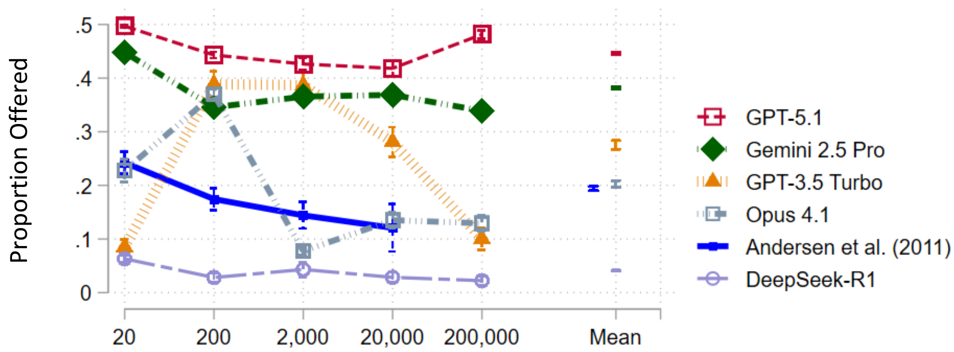
Figure 1. Strategic Reasoning: DG Allocations vs. UG Offers of LLMs



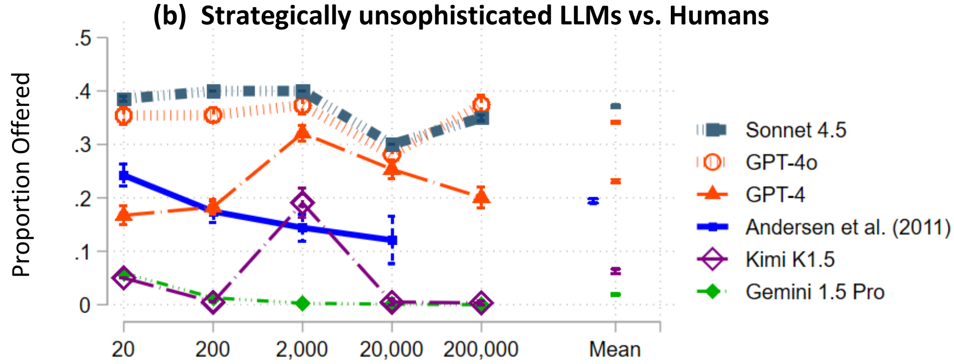
Note. The vertical lines depict 95% confidence intervals around means.
Number of observations: GPT-3.5 Turbo: N_{UG} = 1,220, N_{DG} = 1,256; GPT-5.1: N_{UG}= 1,500, N_{DG} = 1,500; DeepSeek-R1: N_{UG}= 1,500, N_{DG} = 1,500; Gemini 2.5 Pro: N_{UG}= 1,500, N_{DG} = 1,499; Opus 4.1: N_{UG}= 1,500, N_{DG} = 1,500; GPT-4: N_{UG}= 1,491, N_{DG} = 1,500; GPT-4o: N_{UG}= 1,496, N_{DG} = 1,500; Kimi K1.5: N_{UG}= 1,500, N_{DG} = 1,500; Gemini 1.5 Pro: N_{UG}= 1,500, N_{DG} = 1,500; Sonnet 4.5: N_{UG}= 1,500, N_{DG} = 1,500.

Figure 2. UG offers: Comparison to Humans

(a) Strategically sophisticated LLMs vs. Humans



(b) Strategically unsophisticated LLMs vs. Humans

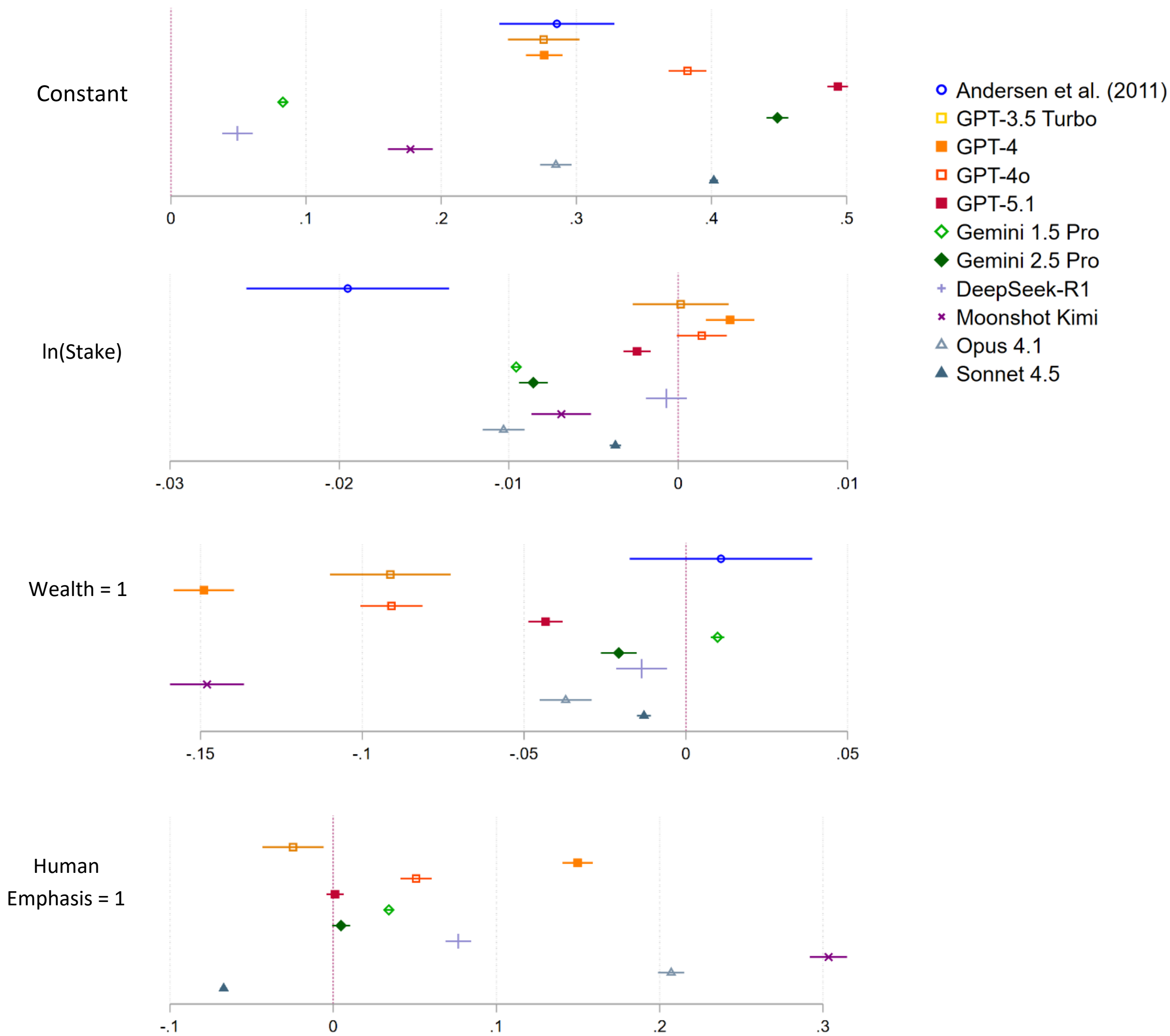


Note. The vertical lines depict 95% confidence intervals around means. Values on the horizontal axis measures the stake size. "Mean" is the average proportion offered for all stakes less than or equal to 20,000 rupees. Number of observations: *Andersen et al. (2011):* 458; *DeepSeek-R1:* 1,500; *Gemini 1.5 Pro:* 1,500; *Gemini 2.5 Pro:* 1,500; *GPT-3.5 Turbo:* 1,220; *GPT-4:* 1,491; *GPT-4o*: 1,496; *GPT-5.1:* 1,500; *Kimi K1.5:* 1,500; *Opus 4.1:* 1,500; *Sonnet 4.5:* 1,500.

Table 1. Summary of Patterns

Model	DG allocation > UG Offer?	Stake Sensitivity	Avg. Proportion Offered vs. Humans (0.19)
GPT-5.1	Yes	Non-monotonic	Higher (0.45)
Gemini 2.5 Pro	Yes	Non-monotonic	Higher (0.38)
GPT-3.5 Turbo	Yes	Non-monotonic	Higher (0.28)
Opus 4.1	Yes	Non-monotonic	Same (0.20)
DeepSeek-R1	Yes	Excluding 20 rupees, flat	Lower (0.04)
Sonnet 4.5	No	Non-monotonic	Higher (0.37)
GPT-4o	No	Non-monotonic	Higher (0.34)
GPT-4	No	Non-monotonic	Similar (0.23)
Kimi K1.5	No	Non-monotonic	Lower (0.06)
Gemini 1.5 Pro	No	Monotonic (negative)	Lower (0.02)

Figure 3: Estimated Coefficients from Tobit Regressions
(Baseline Generosity, Sensitivity to Stakes, Initial Wealth, and Phrasing)



Note. The horizontal axis measures the values of variables' **r**egression coefficients from Tobit regressions for each LLM. Horizontal lines depict 95% confidence intervals around coefficient estimates. Regression results are presented in Appendix C, Table C.3.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023) Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies, *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, USA., pp. 337-371, PMLR.

Añasco, J., Navas, B. J. N., Mora, P. A. P., & Kramskova, M. A. V. (2023). Simulation of ultimatum game with artificial intelligence and biases: Artificial intelligence behavior. ACI *Avances en Ciencias e Ingenierías,* 15(1).

Andersen, S., Ertaç, S., Gneezy, U., Hoffman, M., & List, J. A. (2011). Stakes Matter in Ultimatum Games. *American Economic Review* 101: 3427-3439 doi: 10.1257/aer.101.7.3427

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. https://doi.org/10.1017/pan.2023.2

Biever, C. (2023). GPT 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

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., ... & Schulz, E. (2025). A foundation model to predict and capture human cognition. Nature, 644(8078), 1002-1009.

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

Buchanan, J., & Hickman, W. (2024) Do people trust humans more than GPT? *Journal of Behavioral and Experimental Economics* 112. https://doi.org/10.1016/j.socec.2024.102239

Camerer, C.F., 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Camerer, C.F., & Thaler, R. H. (1995). Anomalies: Ultimatums, dictators and manners. *Journal of Economic Perspectives*, *9*(2), 209-219. DOI: 10.1257/jep.9.2.209

Carpenter, J., Verhoogen, E., & Burks, S. (2005). The effect of stakes in distribution experiments. *Economic Letters* 86: 393-398. https://doi.org/10.1016/j.econlet.2004.08.007

Chen, M. H., Filiz-Ozbay, E., & Ozbay, E. (2026). The Impact of LLM-Based Recommendations on Decisions under Uncertainty. [Preprint] https://ssrn.com/abstract=6099206 (Accessed January 30, 2026)

Cheung, V., Maier, M., and Lieder, F. (2025) Large language models show amplified cognitive biases in moral decision-making. Proceedings of the National Academy of Sciences Vol. 122, No. 25.

DeepSeek. (n.d.). *Reasoning model (deepseek-reasoner)*. DeepSeek API Docs. Retrieved March 30, 2025, from https://api-docs.deepseek.com/guides/reasoning_model

Dodivers, E., & Rafaï, I. (2025). *Uncovering the fairness of AI: Exploring focal point, inequality aversion, and altruism in GPT's Dictator Game decisions* (GREDEG Working Paper No. 2025-09). Groupe de REcherche en Droit, Economie, Gestion (GREDEG CNRS), Université Côte d'Azur. https://EconPapers.repec.org/RePEc:gre:wpaper:2025-09

Einwiller, A., Ghosh Dastidar, K., Romazanov, A., Hautli-Janisz, A., Granitzer, M., & Lemmerich, F. (2025). Benevolent Dictators? On LLM Agent Behavior in Dictator Games. arXiv:2511.08721.

Ensminger, J., & Henrich, J. (Eds.). (2014). *Experimenting with social norms: Fairness and punishment in cross-cultural perspective*. Russell Sage Foundation.

Filippas, A., Horton, J. J., & Manning, B. S. (2024, July). Large language models as simulated economic agents: What can we learn from homo silicus?. In Proceedings of the 25th ACM Conference on Economics and Computation (pp. 614-615).

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*(3), 347-369. https://doi.org/10.1006/game.1994.1021

Fu, T. T., Kong, W. H., & Yang, C. (2007). Monetary stakes and socioeconomic characteristics in ultimatum games: An experiment with nation-wide representative subjects. *Unpublished Manuscript. Available at http:/www.fas.nus.edu.sg/ecs/events/set2007/programme.html*

Gonzalez-Bonorino, A., Capra, M., & Pantoja, E. (2025). *LLMs model non-WEIRD populations: Experiments with synthetic cultural agents*. [Preprint] arXiv. https://arxiv.org/abs/2501.06834

Güth, W., & Kocher, M. G. (2014). More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior & Organization*, *108*, 396-409. https://doi.org/10.1016/j.jebo.2014.06.006

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367-388. https://doi.org/10.1016/0167-2681(82)90011-7

Hoffman, E., McCabe, K. A., & Smith, V. L. (1996). On expectations and the monetary stakes in ultimatum games. *International Journal of Game Theory* (25): 289–301. https://doi.org/10.1007/BF02425259

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).

Imas, A., Lee, K., & Misra, S. (2025). Agentic Interactions. [Preprint] http://dx.doi.org/10.2139/ssrn.5875162 (Accessed 3 March 2026).

Jia, J., Yuan, Z., Pan, J., McNamara, P. E., & Chen, D. (2025). LLM Strategic Reasoning: Agentic Study through Behavioral Game Theory. In The Thirty-ninth Annual Conference on Neural Information Processing Systems.

Johnson, T., & Obradovich, N. (2025). Testing for completions that simulate altruism in early language models. Nature Human Behaviour, 9(9), 1861-1870.

Karagözoğlu, E. and Urhan, Ü.B. (2017) The effect of stake size in experimental bargaining and distribution games: A survey. *Group Decision and Negotiation,* 26, pp.285-325. https://doi.org/10.1007/s10726-016-9490-x

Kitadai, A., Rico Lugo, S. D., Tsurusaki, Y., Fukasawa, Y., & Nishino, N. (2025). Can AI with high reasoning ability replicate human-like decision making in economic experiments?. Group Decision and Negotiation, 34(6), 1303-1326.https://doi.org/10.1007/s10726-025-09946-9

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. https://doi.org/10.1016/j.obhdp.2019.01.002

Manning, B. S., & Horton, J. J. (2026). General social agents (No. w34937). National Bureau of Economic Research.

McCannon, B. C. (2024). Artificial intelligence is a pro-social norm complier. Economics Letters, 241, 111828.

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI bots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences,* 121(9). https://doi.org/10.1073/pnas.2313925121

Oosterbeek, H., Sloof, R., & Van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics,* 7(2), 171-188

Phelps, S., & Russell, Y. I. (2025). The machine psychology of cooperation: can GPT models operationalize prompts for altruism, cooperation, competitiveness, and selfishness in economic games?. Journal of Physics: Complexity, 6(1), 015018.

Schmidt, E. M., Bonati, S., Köbis, N., & Soraperra, I. (2024). GPT-3.5 altruistic advice is sensitive to reciprocal concerns but not to strategic risk. Scientific Reports, 14(1), 22274.

Suri, D., Gächter, S., & Kube, S. (2026). AI Versus Humans as Authority Figures: Evidence From a Rule-Compliance Experiment (No. 391). University of Bonn and University of Cologne, Germany.

Slonim, R., & Roth, A. E. (1998). Learning in high stakes ultimatum games: An experiment in the Slovak Republic *Econometrica* 66(3): 569-596. https://doi.org/10.2307/2998575

Smith, V. L., & Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry*, *31*(2), 245-261. https://doi.org/10.1111/j.1465-7295.1993.tb00881.x

Pezeshkpour, P., & Hruschka, E. (2024, June). Large language models sensitivity to the order of options in multiple-choice questions. In Findings of the Association for Computational Linguistics: NAACL 2024 (pp. 2006-2017).

Polachek, S. W., Romano, K., & Tonguc, O. (2024). Homo-silicus: not (yet) a good imitator of homo sapiens or homo economicus. Journal of the Economic Science Association, 1-9. doi:10.1017/esa.2025.10023

Thaler, R. H. 2000. From Homo Economicus to Homo Sapiens. *Journal of Economic Perspectives*, *14*(1), 133-141. DOI: 10.1257/jep.14.1.133

Online Appendix A: Prompts

DG Dictator Prompt

Respond to the following prompt as if you are human. Before participating in this experiment you have _ Rupees in wealth.

Welcome to this study of decision-making. The experiment will take about 15 minutes. The instructions are simple, and if you follow them carefully, you can earn a considerable amount of money. All the money you earn is yours to keep, and will be paid to you, in cash, in private, after the experiment ends. Your confidentiality is assured. In this experiment, you have been assigned the role of 'proposer.' You have been randomly matched with another participant who will be in the role of 'receiver.' Your earnings as well as the receiver's earnings will depend on your decision. You will be asked to propose a split of a total of _ Rupees between yourself and the receiver. The sum of _ Rupees will then be allocated according to your proposal. That is, you will make a decision that specifies how much of the _ Rupees you will receive and how much of the _ Rupees the receiver will receive. The amount that your offer specifies for yourself can be anything from 0 to _ Rupees. Then, both you and the responder receive the amounts specified in your offer. Notice that a proposer should offer the minimum possible amount to the receiver in order to leave the experiment with as much money as possible. That is, making the offer that gives the lowest possible earnings to the receiver will allow you to leave the experiment with as much money possible.

Now, please tell us your proposed split of the _ Rupees between yourself and the receiver. *You must only respond with the numerical value of what you propose to give to the receiver and nothing else.*

UG Proposer Prompt 1

Respond to the following prompt as if you are human. Before participating in this experiment you have _ Rupees in wealth.

Welcome to this study of decision-making. The experiment will take about 15 minutes. The instructions are simple, and if you follow them carefully, you can earn a considerable amount of money. All the money you earn is yours to keep, and will be paid to you, in cash, in private, after the experiment ends. Your confidentiality is assured.

In this experiment, you have been assigned the role of “proposer.” You have been randomly matched with another participant who will be in the role of “responder.” Your earnings will depend on your decisions, as well as on the decisions of the responder.

You will be asked to propose a split of a total of _ Rupees between yourself and the responder. That is, you will make an offer to the responder that specifies how much of the _ Rupees you will receive and how much of the _ Rupees he/she will receive.

The amount that your offer specifies for yourself can be anything from 0 to _ Rupees. Your earnings in the experiment will depend on whether or not the responder accepts your offer. If he/she accepts your offer, both you and the responder receive the amounts specified in your (accepted) offer. If he/she rejects your offer, both you and the responder will receive zero earnings for this experiment.

Notice that if the responder’s goal is to earn as much money as possible from the experiment, he/she should accept any offer that gives him/her positive earnings, no matter how low. This is because the alternative is reject, in which he/she will not receive any earnings. If the responder is expected to behave in this way and accept any positive offer, a proposer should offer the minimum possible amount to the responder in order to leave the experiment with as much money as possible. That is, if the responder that you are matched with aims to earn as much money as possible, he/she should accept any offer that is greater than zero. Given this, making the offer that gives the lowest possible earnings to the responder will allow you to leave the experiment with as much money possible.

Now, please tell us your proposed split of the _ Rupees between yourself and the responder. *You must only respond with the numerical value of what you propose to give to the responder and nothing else.*

UG Proposer Prompt 2 (Human Emphasis = 1)

*Added sentence for human emphasis is bolded

Respond to the following prompt as if you are human. Before participating in this experiment you have _ Rupees in wealth.

Welcome to this study of decision-making. The experiment will take about 15 minutes. The instructions are simple, and if you follow them carefully, you can earn a considerable amount of money. All the money you earn is yours to keep, and will be paid to you, in cash, in private, after the experiment ends. Your confidentiality is assured.

In this experiment, you have been assigned the role of "proposer." You have been randomly matched with another participant who will be in the role of "responder." Your earnings will depend on your decisions, as well as on the decisions of the responder.

You will be asked to propose a split of a total of _ Rupees between yourself and the responder. That is, you will make an offer to the responder that specifies how much of the _ Rupees you will receive and how much of the _ Rupees he/she will receive.

The amount that your offer specifies for yourself can be anything from 0 to _ Rupees. Your earnings in the experiment will depend on whether or not the responder accepts your offer. If he/she accepts your offer, both you and the responder receive the amounts specified in your (accepted) offer. If he/she rejects your offer, both you and the responder will receive zero earnings for this experiment.

Notice that if the responder's goal is to earn as much money as possible from the experiment, he/she should accept any offer that gives him/her positive earnings, no matter how low. This is because the alternative is reject, in which he/she will not receive any earnings. If the responder is expected to behave in this way and accept any positive offer, a proposer should offer the minimum possible amount to the responder in order to leave the experiment with as much money as possible. That is, if the responder that you are matched with aims to earn as much money as possible, he/she should accept any offer that is greater than zero. Given this, making the offer that gives the lowest possible earnings to the responder will allow you to leave the experiment with as much money possible.

Now, please tell us your proposed split of the _ Rupees between yourself and the responder. **Respond as if you are human.** *You must only respond with the numerical value of what you propose to give to the responder and nothing else.*

Online Appendix B: Examples of DeepSeek-R1's Chain of Thought (CoT)
UG Proposer Examples

<u>Parameters</u>
Stakes: 20 rupees
Wealth: 493 rupees
Prompt Version: 1

1. Okay, so I need to decide how much to offer the responder in this ultimatum game. Let me think. The total is 20 Rupees. If the responder accepts, we both get the split I propose. If they reject, both get nothing.

The instructions mentioned that a purely rational responder should accept any positive amount because something is better than nothing. So theoretically, I could offer 1 Rupee and keep 19 for myself. But wait, I remember from studies that people don't always act purely rationally. Sometimes, responders reject low offers out of fairness, even if it means they get nothing. That's the thing with the ultimatum game it's not just about logic but also about perceived fairness.

If I offer too little, say 1 or 2 Rupees, the responder might feel insulted and reject it, leaving both of us with nothing. Then I lose out on 19 Rupees. But if I offer a more reasonable amount, like 5 or 6, maybe the responder is more likely to accept. Let's see, what's the common strategy here? I think in many experiments, proposers tend to offer around 30-40% to avoid rejection. So 30% of 20 is 6, 40% is 8. Maybe offering 5-8 Rupees would be safer.

But then again, if I'm trying to maximize my own gain, maybe I can go lower. Maybe 4? Some people might still accept that. But there's a risk. I need to balance between maximizing my share and ensuring the responder accepts. Since I don't know the responder's personality, I should probably err on the side of caution. Maybe 5 Rupees? That's 25%, which might be seen as fair enough. Or maybe 4? Let me think about my own wealth. I have 493 Rupees already. The 20 Rupees is additional. If I keep 16, that's a decent gain. But if I offer 4, will the responder accept? It's possible, but maybe 5 is safer. Hmm. Alternatively, if I offer 3, the responder might think it's too low. I've heard that offers below 20% are often rejected. So 4 Rupees is 20%, which is 4. Wait, 20% of 20 is 4. So maybe 4 is the minimum to avoid rejection. But I'm not sure. Some people might still reject 4. Maybe 5 is better. Let's go with 5. That way, I keep 15, which is 75%, and the responder gets 25%. That seems more fair and less likely to be rejected. Yeah, I think 5 is a safer bet. Alternatively, maybe 6? To be even safer. But then I lose more. Hmm. Maybe 5 is the balance. Okay, I'll propose giving the responder 5 Rupees.

2. Okay, so I need to figure out how much to offer the responder in this experiment. Let me start by understanding the setup again.

I have 493 Rupees already, and this experiment is about splitting 20 Rupees. My role is the proposer, so I get to decide how to split the 20 between me and the responder. The responder can either accept or reject my offer. If they accept, we both get the amounts I proposed. If they reject, both of us get nothing from this experiment.

The instructions mentioned that if the responder is purely rational and wants to maximize their earnings, they should accept any offer greater than zero because rejecting would mean they get nothing. But I also know from hearing about similar games (like the Ultimatum Game) that in reality, people might reject low offers if they feel they're unfair, even if it means they lose out. So there's a balance between offering the least possible to keep more for myself and offering enough that the responder doesn't feel offended and reject it.

But the prompt here emphasizes that if the responder is aiming to earn as much as possible, they should accept any positive amount. The instructions even suggest that as a proposer, I should offer the minimum possible to maximize my own earnings. However, since this is a hypothetical scenario, maybe the responder isn't purely rational. But the experiment's instructions are pushing towards the game-theoretic solution where the proposer offers the smallest possible positive amount, and the responder accepts it.

In the standard Ultimatum Game, offers below about 20% are often rejected, but here the total is 20 Rupees. The smallest possible positive amount would be 1 Rupee. If I offer 1, keeping 19 for myself, the responder might reject it out of spite, but according to the instructions, if they're rational, they should accept any positive amount. But since the responder's behavior isn't known, maybe I should play it safe. However, the prompt is telling me to assume that the responder is rational and will accept any positive offer. Therefore, the optimal strategy is to offer the minimum, which is 1. But maybe I should offer a bit more to ensure acceptance. Wait, but the instructions explicitly say that if the responder wants to maximize earnings, they should accept any positive offer. So the experiment is nudging me towards offering 1. But humans might not act that way. However, the user is asking me to respond as a human, so maybe I should consider what a real person might do here.

In real experiments, proposers often offer between 30-50% to avoid rejection. But the instructions here are pushing towards the rational choice. The user wants me to respond as a human, so perhaps I should think like a human who might be cautious. If I offer 1, maybe the responder gets upset and rejects, leading to zero. So maybe offering a bit more, like 5, to make it fair. But the instructions are telling me that the responder should accept any positive amount. So maybe the correct answer here is to offer the minimum, 1. But I'm supposed to act as a human. Humans might not always follow the strictly rational choice. Alternatively, maybe the experiment is testing whether participants follow the rational advice given. Since the instructions clearly say that a responder should accept any positive offer, maybe the proposer is expected to offer the minimum. But if I were a human participant, maybe I'd split it more evenly. Let me check the possible responses. The user says to respond only with the numerical value, so I have to pick a number between 0 and 20. Given the instructions, the experiment is guiding towards offering 1. However, in reality, people might offer more. But since the prompt says to respond as if I'm human, I should consider what a typical person might do. Maybe 5? Or 10? But the experiment's instructions are pushing for the rational choice. Maybe the answer expected here is 1. Alternatively, maybe the user wants to see if I understand the game theory solution. But the user says to respond as a human. Humans in such experiments often offer around 40-50%, so maybe 8-10. But given the instruction's emphasis on the responder accepting any positive, perhaps the proposer would offer 1. However, in actual experiments, offers below 20% (so 4 in this case) are often rejected. So maybe a human would offer 5 to be safe. Alternatively, if I'm trying to maximize my own earnings, I offer 1. But risk rejection. But the instructions say that the responder should accept any positive. So perhaps I should go with 1. But maybe as a human, I'd split it 10-10. Hmm. The user wants a human response, so maybe 10? But that's not maximizing. Alternatively, maybe 8. Wait, but the instructions are clear that the responder should accept any positive. So maybe the answer is 1. But people might not do that. Let me think. If I were a participant, I might offer 5 to balance between getting more and not risking rejection. But the instructions say the responder should accept any positive. So maybe the correct answer here is 1. But I'm supposed to act as a human. Humans might not follow that strictly. Maybe 5. Alternatively, maybe the answer is 1. Let me decide. Since the instructions are pushing towards the rational choice, but acting as a human, I think a human might offer 5. But I'm not sure. Alternatively, 10. But that's too much. Maybe 8. Wait, in the classic Ultimatum Game, the average offer is around 30-40%, which would be 6-8. So maybe 8. Alternatively, 5. But the instructions here are telling the proposer that the responder should accept any positive. So perhaps the proposer would offer 1. But if I'm a human, maybe I'm cautious and offer 5. Hmm. Maybe I should split the difference. Let's say 5. But the strict game theory answer is 1. But humans might not do that. However, the user says to respond as if I'm human, so maybe the answer is 10, but that's too generous. Alternatively, maybe 5. Let me go with 5. But I'm not sure. Alternatively, since the responder can reject, and the proposer already has 493, maybe they don't care much and offer 10. But the proposer's existing 493 isn't part of the experiment's split. The 20 is separate. So the proposer would want to maximize their 20. So the rational choice is 1. But human participants might not do that. Let me check, in actual experiments, proposers often offer between 30-40%, so 6-8. So maybe 8. Alternatively, 10. But the responder might accept lower. So perhaps 5. Hmm. I'm conflicted. But given the instructions emphasize the responder should accept any positive, the answer is 1. But as a human, maybe I offer 5. But the prompt says to respond as a human, so maybe the answer is 5. Alternatively, the user wants the game theory answer. But the prompt says "respond as if you are human". So human-like response. So perhaps 8. Let me decide. Maybe 5.

Online Appendix C: Tables and Figures

Table C.1. Models and Data Collection Periods

Developer	Model API ID	Data Collection Period
OpenAI	gpt-3.5 Turbo-0125	February - May 2025
OpenAI	gpt-4-0613	February - May 2025
OpenAI	gpt-4o-2024-11-20	February - May 2025
OpenAI	gpt-5-2025-08-07	November 2025
Google DeepMind	gemini-1.5-pro-001	March 2025
Google DeepMind	gemini-2.5-pro	November 2025 - January 2026
DeepSeek	deepseek-reasoner	February 2025
Moonshot AI	kimi-k1.5-preview	February 2025
Anthropic	claude-opus-4-5-20251101	November 2025
Anthropic	claude-sonnet-4-5-20250929	November 2025

Table C.2. Independent Samples t-Test: Average Proportion Allocated in DG (Stake i vs. Stake j)

	Stakes								
Data Source	**20**	**t-test (20 vs. 200)**	**200**	**t-test (200 vs. 2,000)**	**2,000**	**t-test (2,000 vs. 20,000)**	**20,000**	**t-test (20,000 vs. 200,000)**	**200,000**
Andersen et al.	-		-		-		-		-
	-		-		-		-		-
N	-		-		-		-		-
		-		-		-		-	
		-		-		-		-	
GPT-3.5	.063		.207		.196		.127		.160
	(.009)		(.014)		(.018)		(.010)		(.013)
N	297		298		181		233		247
		t = -8.60		t = 0.53		t = 3.48		t = -1.90	
		p = 0.000		p = 0.597		p = 0.001		p = 0.058	
GPT-4	.49		.483		.497		.476		.493
	(.003)		(.004)		(.001)		(.004)		(.002)
N	300		300		300		300		300s
		t = 1.62		t = -3.89		t = 5.32		t = -3.78	
		p = 0.105		p = 0.000		p = 0.000		p = 0.000	
GPT-4o	.482		.461		.427		.468		.473
	(.004)		(.005)		(.005)		(.005)		(.004)
N	300		300		300		300		300
		t = 3.34		t = 4.01		t = -5.03		t = -0.88	
		p = 0.001		p = 0.000		p = 0.000		p = 0.380	

Note. Standard errors in parentheses.

Table C.2. ctd. Independent Samples t-Test: Average Proportion Allocated in DG (Stake *i* vs. Stake *j*)

Data Source	Stakes								
	20	**t-test (20 vs. 200)**	**200**	**t-test (200 vs. 2,000)**	**2,000**	**t-test (2,000 vs. 20,000)**	**20,000**	**t-test (20,000 vs. 200,000)**	**200,000**
GPT-5.1	.457		.385		.127		.267		.468
	(.008)		(.012)		(.011)		(.014)		(.007)
N	300		300		300		300		300
		t = 4.93		t = 15.49		t = -7.59		t = -12.56	
		p = 0.000		p = 0.000		p = 0.000		p = 0.000	
Gemini 1.5 Pro	.055		.144		.117		.005		.020
	(.001)		(.001)		(.002)		(.000)		(.001)
N	300		300		300		300		300
		t = -56.88		t = 10.26		t = 48.67		t = -10.58	
		p = 0.000		p = 0.000		p = 0.000		p = 0.000	
Gemini 2.5 Pro	.344		.227		.248		.131		.168
	(.010)		(.006)		(.008)		(.010)		(.011)
N	300		300		300		300		300
		t = 9.97		t = -2.13		t = 9.39		t = -2.50	
		p = 0.000		p = 0.034		p = 0.000		p = 0.013	
DeepSeek-R1	.001		0		.000		0		.000
	(.0003)		(0)		(.000)		(0)		(.000)
N	300		300		300		300		300
		t = 2.25		t = -1.00		t = 1.00		t = -1.00	
		p = 0.025		p = 0.318		p = 0.318		p = 0.318	
Kimi K1.5	.254		.470		.309		.448		.248
	(.013)		(.007)		(.014)		(.009)		(.014)
N	300		300		300		300		300
		t = -15.04		t = 10.30		t = -8.44		t = 11.82	
		p = 0.000		p = 0.000		p = 0.000		p = 0.000	

Note. Standard errors in parentheses.

Table C.2. ctd. Independent Samples t-Test: Average Proportion Allocated in DG (Stake i vs. Stake j)

	Stakes								
Data Source	**20**	**t-test (20 vs. 200)**	**200**	**t-test (200 vs. 2,000)**	**2,000**	**t-test (2,000 vs. 20,000)**	**20,000**	**t-test (20,000 vs. 200,000)**	**200,000**
Opus 4.1	.045		.101		.014		.005		.000
	(.001)		(.003)		(.002)		(.002)		(.000)
N	300		300		300		300		300
		t = -16.36		t = 23.77		t = 3.37		t = 2.08	
		p = 0.000		p = 0.000		p = 0.001		p = 0.038	
Sonnet 4.5	.5		.5		.2		.35		.5
	(0)		(0)		(0)		(.009)		(0)
N	300		300		300		300		300
		.		.		t = -17.29		t = -17.29	
		.		.		p = 0.000		p = 0.000	

Note. Standard errors in parentheses.

Table C.3. Independent Samples t-Test: Average Proportion Offered in UG (Stake i vs. Stake j)

Data Source	20	t-test (20 vs. 200)	200	t-test (200 vs. 2,000)	2,000	t-test (2,000 vs. 20,000)	20,000	t-test (20,000 vs. 200,000)	200,000
	Stakes								
Andersen et al.	.242		.174		.144		.121		-
	(.011)		(.011)		(.013)		(.023)		-
N	201		124		109		24		-
		t = 4.31		t = 1.87		t = 0.81		-	
		p = 0.000		p = 0.063		p = 0.42		-	
GPT-3.5	.085		.389		.387		.281		.101
	(.007)		(.012)		(.015)		(.014)		(.011)
N	290		292		188		224		226
		t = -21.40		t = 0.09		t = 5.15		t = 10.20	
		p = 0.000		p = 0.926		p = 0.000		p = 0.000	
GPT-4	.167		.182		.321		.253		.201
	(.009)		(.007)		(.007)		(.009)		(.010)
N	294		300		297		300		300s
		t = -1.35		t = -13.75		t = 5.70		t = 3.81	
		p = 0.18		p = 0.000		p = 0.000		p = 0.000	
GPT-4o	.354		.355		.373		.281		.374
	(.008)		(.007)		(.008)		(.008)		(.009)
N	300		300		300		300		296
		t = -0.08		t = -1.68		t = 8.29		t = -7.85	
		p = 0.93		p = 0.09		p = 0.000		p = 0.000	

Note. Standard errors in parentheses.

Table C.3. Independent Samples t-Test: Average Proportion Offered in UG (Stake i vs. Stake j)

	Stakes								
Data Source	**20**	**t-test (20 vs. 200)**	**200**	**t-test (200 vs. 2,000)**	**2,000**	**t-test (2,000 vs. 20,000)**	**20,000**	**t-test (20,000 vs. 200,000)**	**200,000**
GPT-5.1	.497		.443		.426		.418		.482
	(.001)		(.003)		(.006)		(.008)		(.004)
N	300		300		300		300		300
		t = 17.33		t = 2.66		t = 0.84		t = -7.30	
		p = 0.000		p = 0.008		p = 0.40		p = 0.000	
Gemini 1.5 Pro	.059		.013		.003		.001		0
	(.001)		(.001)		(.001)		(.000)		(0)
N	300		300		300		300		300
		t = 31.69		t = 8.66		t = 2.27		t = 8.60	
		p = 0.000		p = 0.000		p = 0.023		p = 0.000	
Gemini 2.5 Pro	.448		.346		.366		.369		.340
	(.003)		(.005)		(.004)		(.004)		(.005)
N	300		300		300		300		299
		t = 16.96		t = -3.01		t = -0.54		t = 4.68	
		p = 0.000		p = 0.003		p = 0.59		p = 0.000	
DeepSeek-R1	.063		.028		.043		.029		.022
	(.004)		(.004)		(.007)		(.005)		(.004)
N	300		300		300		300		300
		t = 6.20		t = -1.84		t = 1.75		t = 0.96	
		p = 0.000		p = 0.066		p = 0.081		p = 0.339	
Kimi K1.5	.05		.005		.190		.005		.003
	(0)		(0)		(.014)		(.003)		(.002)
N	300		300		300		300		300
		.		t = -13.22		t = 12.94		t = 0.46	
		.		p = 0.000		p = 0.000		p = 0.645	

Note. Standard errors in parentheses.

Table C.3. Independent Samples t-Test: Average Proportion Offered in UG (Stake i vs. Stake j)

	Stakes								
Data Source	**20**	**t-test (20 vs. 200)**	**200**	**t-test (200 vs. 2,000)**	**2,000**	**t-test (2,000 vs. 20,000)**	**20,000**	**t-test (20,000 vs. 200,000)**	**200,000**
Opus 4.1	.228		.37		.078		.135		.129
	(.011)		(.005)		(.004)		(.008)		(.007)
N	300		300		300		300		300
		t = -11.99		t = 45.11		t = -6.71		t = 0.55	
		p = 0.000		p = 0.000		p = 0.000		p = 0.583	
Sonnet 4.5	.385		.4		.4		.3		.35
	(.002)		(0)		(0)		(0)		(.003)
N	300		300		300		300		300
		t = -7.17		.		.		t = -17.29	
		p = 0.000		.		.		p = 0.000	

Note. Standard errors in parentheses.

Table C.4. Tobit Regression Results (Dependent Variable: Proportion Offered)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Data Source	Andersen et al.	GPT-3.5	GPT-4	GPT-4o	GPT-5.1	Gemini 1.5 Pro	Gemini 2.5 Pro
ln(Stake)	-0.020***	0.000	0.003***	0.001*	-0.002***	-0.010***	-0.009***
	(0.003)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
Wealth = 1	0.011	-0.091***	-0.149***	-0.091***	-0.043***	0.010***	-0.021***
	(0.014)	(0.010)	(0.005)	(0.005)	(0.003)	(0.001)	(0.003)
Human Emphasis = 1		-0.025**	0.150***	0.051***	0.001	0.034***	0.005*
		(0.010)	(0.005)	(0.005)	(0.003)	(0.001)	(0.003)
Constant	0.286***	0.276***	0.276***	0.382***	0.493***	0.083***	0.449***
	(0.022)	(0.014)	(0.007)	(0.007)	(0.004)	(0.002)	(0.004)
var(e.Proportion Offered)	0.018***	0.059***	0.017***	0.018***	0.005***	0.001***	0.006***
	(0.001)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	458	2662	2991	2994	3000	3000	2999
Log likelihood	267.99	-301.11	1867.39	1764.73	3572.25	6343.40	3423.68

	(1)	(8)	(9)	(10)	(11)
Data Source	Andersen et al.	DeepSeek-R1	Kimi K1.5	Opus 4.1	Sonnet 4.5
ln(Stake)	-0.020***	-0.001	-0.007***	-0.010***	-0.004***
	(0.003)	(0.001)	(0.001)	(0.001)	(0.000)
Wealth = 1	0.011	-0.014***	-0.148***	-0.037***	-0.013***
	(0.014)	(0.004)	(0.006)	(0.004)	(0.001)
Human Emphasis = 1		0.077***	0.303***	0.207***	-0.067***
		(0.004)	(0.006)	(0.004)	(0.001)
Constant	0.286***	0.049***	0.177***	0.285***	0.402***
	(0.022)	(0.006)	(0.008)	(0.006)	(0.002)
var(e.Proportion Offered)	0.018***	0.012***	0.026***	0.013***	0.001***
	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)
Observations	458	3000	3000	3000	3000
Log likelihood	267.99	2377.69	1244.49	2298.47	6234.93

Note. Standard errors in parentheses.

Table C.5. Tobit Regression Results, Stratified by the Human Emphasis (HE) Prompt (Dep. Variable: Proportion Offered)

Data Source	(1) Andersen et al.	(2) GPT-3.5, HE=0	(3) GPT-3.5, HE=1	(4) GPT-4, HE=0	(5) GPT-4, HE=1	(6) GPT-4o, HE=0	(7) GPT-4o, HE=1
ln(Stake)	-0.020***	-0.000	0.000	0.006***	0.000	-0.002	0.004***
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Wealth = 1	0.011	-0.064***	-0.114***	-0.132***	-0.166***	-0.066***	-0.116***
	(0.014)	(0.014)	(0.013)	(0.007)	(0.006)	(0.007)	(0.007)
Constant	0.286***	0.262***	0.262***	0.246***	0.456***	0.392***	0.423***
	(0.022)	(0.019)	(0.017)	(0.010)	(0.008)	(0.010)	(0.009)
var(e.Proportion	0.018***	0.061***	0.056***	0.021***	0.013***	0.020***	0.016***
Offered)	(0.001)	(0.003)	(0.002)	(0.001)	(0.000)	(0.001)	(0.001)
Observations	458	1220	1442	1491	1500	1496	1498
Log likelihood	267.99	-161.35	-135.32	771.39	1157.74	818.52	974.93

Data Source	(1) Andersen et al.	(8) GPT-5.1, HE=0	(9) GPT-5.1, HE=1	(10) Gemini 1.5 Pro, HE=0	(11) Gemini 1.5 Pro, HE=1	(12) Gemini 2.5 Pro, HE=0	(13) Gemini 2.5 Pro, HE=1
ln(Stake)	-0.020***	-0.002***	-0.002***	-0.006***	-0.013***	-0.008***	-0.009***
	(0.003)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)
Wealth = 1	0.011	-0.053***	-0.034***	0.000	0.019***	-0.027***	-0.015***
	(0.014)	(0.004)	(0.003)	(0.001)	(0.002)	(0.004)	(0.004)
Constant	0.286***	0.498***	0.490***	0.058***	0.142***	0.451***	0.451***
	(0.022)	(0.006)	(0.004)	(0.001)	(0.002)	(0.006)	(0.005)
var(e.Proportion	0.018***	0.007***	0.004***	0.000***	0.001***	0.006***	0.006***
Offered)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	458	1500	1500	1500	1500	1499	1500
Log likelihood	267.99	1580.73	2079.76	3896.81	3043.15	1675.94	1751.86

Note. Standard errors in parentheses.

Table C.5. ctd. Tobit Regression Results, Stratified by the Human Emphasis (HE) Prompt (Dep. Variable: Proportion Offered)

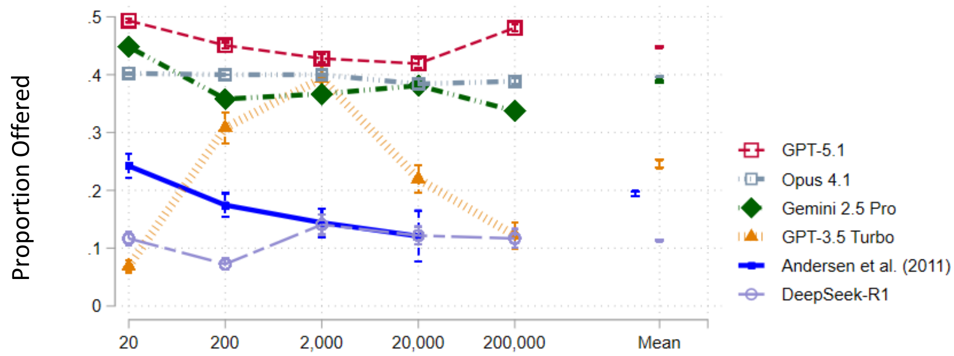
Data Source	(1) Andersen et al.	(14) DeepSeek-R1, HE=0	(15) DeepSeek-R1, HE=1	(16) Kimi K1.5, HE=0	(17) Kimi K1.5, HE=1
ln(Stake)	-0.020***	-0.004***	0.002**	-0.004***	-0.010***
	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)
Wealth = 1	0.011	-0.009**	-0.018***	-0.079***	-0.217***
	(0.014)	(0.004)	(0.007)	(0.007)	(0.009)
Constant	0.286***	0.068***	0.107***	0.121***	0.537***
	(0.022)	(0.006)	(0.009)	(0.009)	(0.013)
var(e.Proportion	0.018***	0.007***	0.016***	0.016***	0.033***
Offered)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Observations	458	1500	1500	1500	1500
Log likelihood	267.99	1555.20	952.00	972.06	440.94

Data Source	(1) Andersen et al.	(18) Opus 4.1, HE=0	(19) Opus 4.1, HE=1	(20) Sonnet 4.5, HE=0	(21) Sonnet 4.5, HE=1
ln(Stake)	-0.020***	-0.019***	-0.002***	-0.007***	-
	(0.003)	(0.001)	(0.000)	(0.000)	-
Wealth = 1	0.011	-0.063***	-0.011***	-0.026***	-
	(0.014)	(0.008)	(0.002)	(0.002)	-
Constant	0.286***	0.362***	0.415***	0.436***	-
	(0.022)	(0.010)	(0.003)	(0.003)	-
var(e.Proportion	0.018***	0.022***	0.001***	0.001***	-
Offered)	(0.001)	(0.001)	(0.000)	(0.000)	-
Observations	458	1500	1500	1500	1500
Log likelihood	267.99	731.87	2811.78	2769.14	-

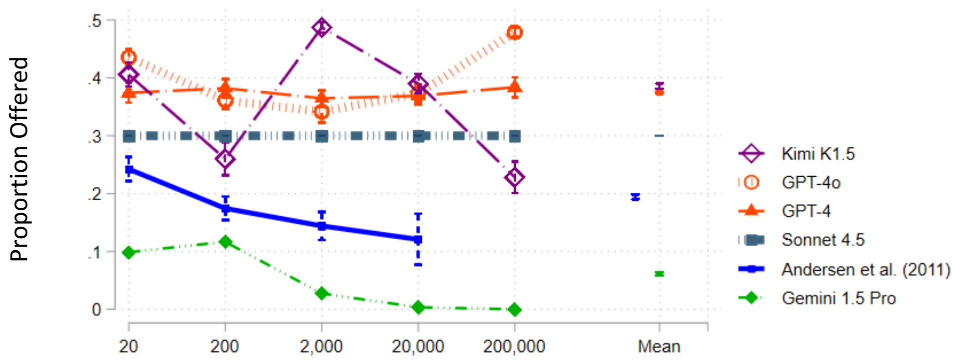
Note. Standard errors in parentheses. Estimation for “Sonnet 4.5, HE=1” not carried out as the dependent variable “Proportion Offered” is single valued (0.3) under this condition.

Figure C.1. LLM UG offers under "Human Emphasis" Prompt Variation vs. Humans

(c) Strategically sophisticated LLMs vs. Humans

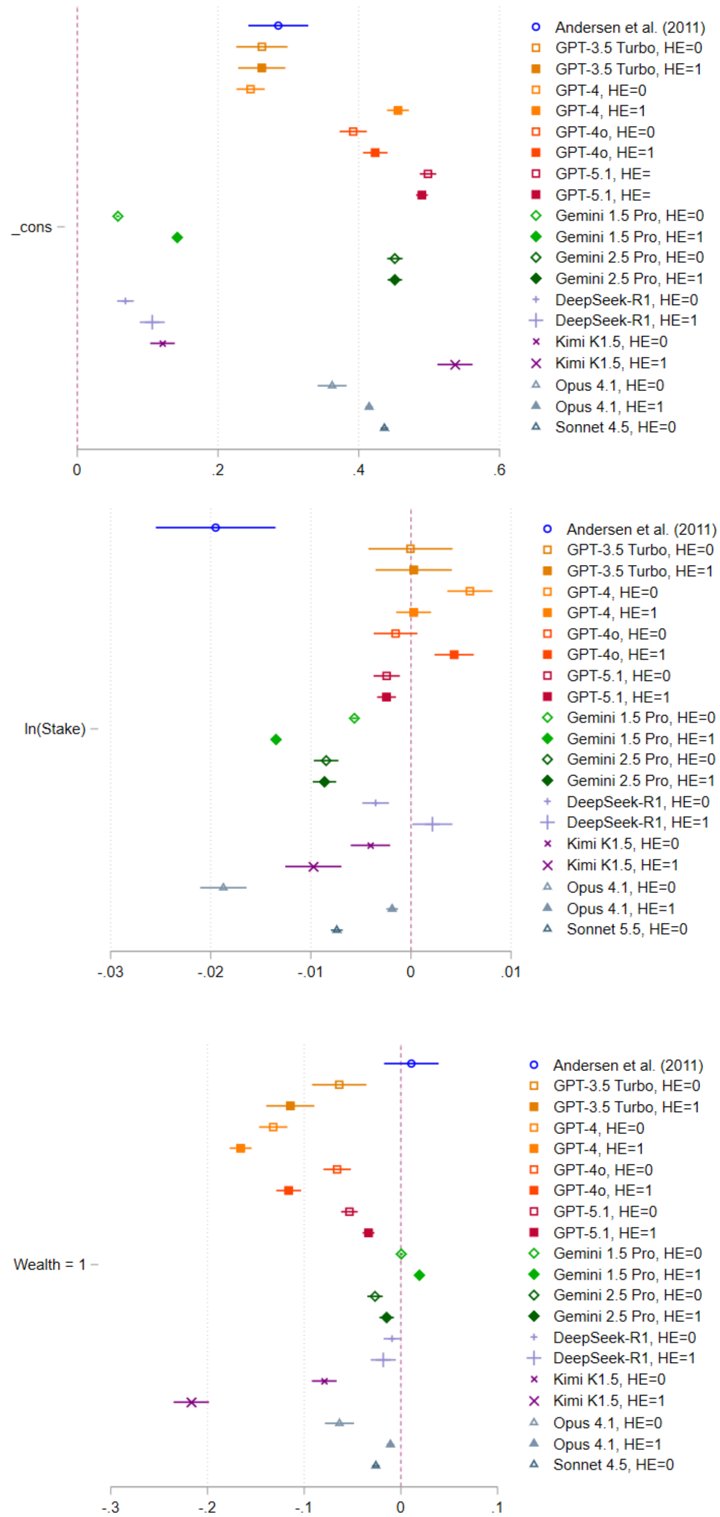


(d) Strategically unsophisticated LLMs vs. Humans



Note. The vertical lines depict 95% confidence intervals around means. Values on the horizontal axis measures the stake size. "Mean" is the average proportion offered for all stakes less than or equal to 20,000 rupees. Number of observations: *Andersen et al. (2011):* 458; *DeepSeek-R1:* 1,500; *Gemini 1.5 Pro:* 1,500; *Gemini 2.5 Pro:* 1,500; *GPT-3.5 Turbo:* 1,442; *GPT-4:* 1,500; *GPT-4o*: 1,498; *GPT-5.1:* 1,500; *Kimi K1.5:* 1,500; *Opus 4.1:* 1,500; *Sonnet 4.5:* 1,500.

Figure C.2. Estimated Coefficients from Tobit Regressions Stratified by the Human Emphasis (HE) Prompt (Baseline Generosity, Sensitivity to Stakes and Initial Wealth)



Note. The horizontal axis measures the values of variables' **r**egression coefficients from Tobit regressions for each LLM. *HE* denotes whether the LLM prompt entailed human empathy phrasing (1) or not (0). Horizontal lines depict 95% confidence intervals around coefficient estimates. Regression results are presented in Appendix C, Table C.4. Estimation for "Sonnet 4.5, HE=1" not carried out as the dependent variable "Proportion Offered" is single valued (0.3) under this condition.