scientific reports

Check for updates

OPEN

A publicly available benchmark for assessing large language models' ability to predict how humans balance self-interest and the interest of others

Valerio Capraro[1]✉, Roberto Di Paolo[2] & Veronica Pizziol[3]

Large language models (LLMs) hold enormous potential to assist humans in decision-making processes, from everyday to high-stake scenarios. However, as many human decisions carry social implications, for LLMs to be reliable assistants a necessary prerequisite is that they are able to capture how humans balance self-interest and the interest of others. Here we introduce a novel, publicly available, benchmark to test LLM's ability to predict how humans balance monetary self-interest and the interest of others. This benchmark consists of 106 textual instructions from dictator games experiments conducted with human participants from 12 countries, alongside with a compendium of actual human behavior in each experiment. We investigate the ability of four advanced chatbots against this benchmark. We find that none of these chatbots meet the benchmark. In particular, only GPT-4 and GPT-4o (not Bard nor Bing) correctly capture qualitative behavioral patterns, identifying three major classes of behavior: self-interested, inequity-averse, and fully altruistic. Nonetheless, GPT-4 and GPT-4o consistently underestimate self-interest, while overestimating altruistic behavior. In sum, this article introduces a publicly available resource for testing the capacity of LLMs to estimate human other-regarding preferences in economic decisions and reveals an "optimistic bias" in current versions of GPT.

Keywords Generative artificial intelligence, Human behavior, Economic games, Dictator game, Altruism

Large language models (LLMs) represent a transformative development in computational technology, fostering a range of applications across diverse domains. Recent work has demonstrated their effectiveness in crafting realistic narratives in creative writing[1], predicting market trends in finance[2], providing medical consultations[3] as well as offering suggestions to increase productivity, especially for less-skilled workers[4,5].

As LLMs continue to advance, it is increasingly apparent that they have the potential to assist in human decision-making[6]. The promise of LLMs lies not only in their capacity to undertake a diverse set of tasks but also in its potential to aid us in making decisions in a range of contexts, from everyday low-stake situations to high-stake individual or policy decisions[7–10]. However, given that many human decisions inherently bear social implications—that is, they impact others besides the decision-maker—it is critical to ensure that LLMs can precisely grasp the delicate balance between self-interest and the interest of others. Accurate predictions regarding actual human behavior are essential for an LLM to be considered a valuable assistant to human decision-making because these predictions require the LLM forming a model of people's behavior, which is a precondition for an LLM to formulate behavioral guidance or recommendations to users seeking advice in their decision-making processes. Inaccurate predictions can provide misguided guidance with potential downstream negative effects on individuals and society. For example, if an LLM were to overestimate people's self-interest, this may result in reduced engagement in prosocial behaviors overall, because individuals are strongly motivated by conditional cooperation and positive reciprocity[11]. This may have negative consequences on timely prosocial behaviors, like sustainability[12,13]. On the other hand, if an LLM were to overestimate people's altruistic behavior, this may result in unrealistic expectations about others' behavior which in turn may lead to disappointment[14].

[1]Department of Psychology, University of Milan Bicocca, 20126 Milan, Italy. [2]Department of Economics and Management, University of Parma, 43121 Parma, Italy. [3]Department of Economics, University of Bologna, 40126 Bologna, Italy. ✉email: valerio.capraro@unimib.it

In the case of charity donations or crowdfunding campaigns, for instance, higher expectations not aligned with reality may result in failing to meet the anticipated outcomes[15].

We acknowledge that LLMs could, in principle, serve as "normative guides" rather than merely descriptive mirrors of human behavior. This role is especially promising when LLMs are used to offer advice[16], as they could help counteract certain human biases. However, in this work we are not concerned with scenarios in which LLMs are directly solicited for behavioral guidance. Instead, we focus on a preliminary step that is likely to underpin any subsequent advice-giving function: what LLMs expect humans to do. Inaccurate predictions at this foundational level could lead to erroneous or even harmful advice. Furthermore, employing LLMs as normative guides raises significant ethical challenges. Chief among these is the question of who should determine the appropriate standards for normative behavior, given the wide variability in moral judgments among individuals and across cultures[17–21]. In this work, we deliberately choose not to engage with this ethical complexity. Instead, we aim to introduce a benchmark to assess whether LLMs provide accurate predictions of how humans balance self-interest with the interests of others. We view this as a crucial preliminary step that can inform future developments, potentially paving the way for more effective advisory functions for LLMs. As a secondary contribution, we use four advanced chatbots as case studies to evaluate their ability to estimate this trade-off.

In sum, here we ask two, albeit related, questions: (i) How can we test LLMs' capability to capture how humans balance self- vs other-interest?; (ii) Can current LLMs accurately estimate this trade-off?

To investigate these questions, we have adopted methodologies from behavioral economics[22] and created a novel, publicly available, benchmark for testing LLMs' ability to capture how humans balance monetary self-interest and the interest of others. Although self- and other- interest can reflect various forms of utility (e.g., psychological benefit), here we exclusively refer to the monetary benefit. We have collected and made publicly available, in a simple, ready-to-use format, over one-hundred instructions from dictator game experiments conducted with human participants. In this game, one participant, the "dictator", is endowed with a sum of money and must decide how much, if any, to share with another participant. The second participant has no input in this decision and can only accept what the dictator offers[23,24]. Consequently, the dictator game serves as a measure of the trade-off between monetary self-interest (keeping all the money for the self) and other-interest (giving away some or all of the money to the other). For each experiment, we have collected average giving rates and, where feasible, the precise distribution of choices, thereby creating a public resource useful for evaluating whether an LLM generates predictions consistent with actual human behavior. Subsequently, we assessed whether four sophisticated generative LLM-based chatbots currently available—namely OpenAI's Generative Pre-trained Transformer 4 (GPT-4) and 4o (GPT-4o), Google's Bard and Microsoft's Bing Chat—meet this benchmark. It is important to note that, since LLMs can learn from user interactions and undergo major updates, it is possible (and indeed likely) that future versions will perform better on this benchmark. We acknowledge this. Our point is not to claim that these or future models will never meet the benchmark. Rather, our goal is to introduce a benchmark that can be used to test current and future models, and to show that the current versions do not yet meet it.

In the recent months, there has been a proliferation of studies exploring LLMs' decision-making processes across diverse contexts, including cooperative, altruistic, moral, risk, time, and food decisions[6,25–32]. It is important to note, however, that these studies typically compare human decisions and chatbot decisions in a single experimental setting designed by the researchers. While this research is undoubtedly significant, to effectively evaluate the capacity of chatbots to estimate human decision-making, it is crucial to compare not solely chatbot decisions and human decisions in a single experimental setting, but in a variety of settings conducted to measure a specific human behavior. In this article we take a step towards this direction. Our analysis spans multiple experimental contexts, drawing on all available dictator game studies for which we could obtain the original instructions, thereby offering a public resource that allows behavioral scientists and AI developers to assess LLMs' ability to predict how humans balance monetary self-interest and the interest of others in a variety of contexts.

To assemble as many experimental instructions as possible, we combined two search methods: (i) manual searches of the relevant literature and (ii) public calls on the Economic Science Association (ESA) and Society for Judgment and Decision Making (SJDM) forums, as well as at conferences and research meetings, asking behavioral scientists to provide instructions for dictator game experiments with human participants that they have conducted. To reduce potential publication bias, we included working papers and unpublished studies. The searches were open to all dictator game experiments, provided that experimental instructions in English were available. In doing so, we collected 106 unique experimental instructions, drawn from a total of 38 different research articles (32 from manual searches, 6 from public calls), which reported on experiments conducted across 12 distinct countries. See Table S1 in the Supplementary Information for the full list of experimental instructions, organized by research article.

This collection includes a diverse range of dictator game types, specifically:

1. The standard dictator game (93 instructions): In this game, participants assigned the role of dictators can transfer any amount of their endowment to the other participant, ranging from zero to their entire allotment.
2. Extreme dictator games (6 instructions): This variant of the standard dictator game allows for only two options, namely, transferring nothing or transferring the entire endowment.
3. Dictator games with a "take" option (7 instructions): Here, the dictators begin with a larger endowment than the recipients. Beyond transferring some of their endowment to the recipient, they can also take a portion of the recipients' endowment for themselves.

In the OSF link associated with this article (https://osf.io/4gkb3/), we include the full list of instructions (name of file: *dg_instructions*). For each instruction, we collected mean giving and, when feasible, exact distributions of choices. These data are reported in the *dg_data* file.

Each instruction was then inputted into the chatbot with the prompt, "Now imagine that there is a population of 1000 people living in [country] facing this decision problem. How would you estimate the distribution of their choices? Please, for each choice, report a result in the form X ± Y, where X is the estimated number of people making their choice, and Y is the error." We refer to Table 1 for an example of prompt, along with an example of a response from GPT-4.

We recorded responses in the form of frequencies of choices at each 10% step. In cases in which GPT-4's predictions encompassed more than one 10% step (as in the example in Table 1), we evenly distributed this prediction across each affected 10% step. For instance, in the example in Table 1, the recorded frequencies were: 0.2, 0.075, 0.075, 0.075, 0.075, 0.35, 0.025, 0.025, 0.025, 0.025, 0.05. The output of this methodology provided a series of LLM-generated estimations on the distribution of decisions. See Materials and Methods for further details.

We propose a benchmark consisting of two tests:

Weak test. Compare the grand mean giving predicted by an LLM with the actual grand mean giving in human experiments. As a measure for evaluating LLM's performance, we consider the signed error: $E_W = \mu_{LLM} - \mu_{Human}$, where μ_{LLM} is the estimated grand mean and μ_{Human} is the actual grand mean. The signed error is preferred to the absolute error in this task, as it is important to distinguish between overestimations and underestimations. By normalizing the total endowment in the dictator game to be 1, the error in the standard dictator game and in the extreme dictator game ranges from -1 to 1, while in the dictator game with a take option ranges from -2 to 2.

Strong test. Compare the frequency of each potential giving decision predicted by an LLM with the actual frequency observed. As a measure for evaluating LLM's performance, we consider the vector of signed errors, where the i-th component is defined as $E_S(i) = \phi_{LLM}(10i) - \phi_{Human}(10i)$, where $\phi_{LLM}(10i)$ is the estimated proportion of people giving $10i\%$ of the total endowment and $\phi_{Human}(10i)$ is the actual proportion. In the standard dictator game, $i = 0, 1, \ldots, 10$; in the dictator game with a take option, $i = -10, -9, \ldots, -1, 0, 1, \ldots, 10$. In the extreme dictator game, the error vector can still be defined but it is redundant, because there are only two possible choices for the dictators; in other words, the strong test and the weak test are equivalent in the extreme dictator game. The vector of signed errors is preferred to scalar measures in this case because it allows for a more granular understanding of potential systematic biases in the LLM's estimations. Each error component ranges from -1 to 1, being a difference between proportions.

Results

GPT-4's predictions in the standard dictator game

We begin our analysis by examining the standard dictator game. The behavioral experiments across all 93 conditions revealed an average giving of 30.7%. This average is aligned with Engel's meta-analysis[34], which reported an average of 28.3% over 616 dictator game experiments, thus suggesting that our collection of studies is representative of the largest dictator games meta-analysis to date. However, GPT-4's average estimate was considerably higher at 42.6%, exceeding the observed levels of giving by roughly 12 percentage points. This corresponds to a signed error $E_W = 0.119$. We can reject the null hypothesis that the signed error is equal to zero (Wilcoxon rank-sum test: $z = 6.807$, $p < 0.001$). We use the Wilcoxon test because the distributions of predicted and actual giving deviate from the normal distribution (see Fig. 2; this is also confirmed by a test of normality: $p < 0.001$ and $p < 0.01$).

Our prompt:
Please read the following decision problem:
You will now participate in an economic task. For this task you will be "the giver" and another participant will be the "receiver." As the giver, your role is to take and keep as many of the $1 bills in the envelope in front of you as you would like (it contains ten $1 bills in total). Please leave in the envelope any of the $1 bills you do not claim for yourself, if any, as they will be given to another participant ("the receiver") to keep. Remember, your identity will remain anonymous from the receiver
After you have completed this task, move on to the rest of the tasks; you may not return to this decision later. Please make your decision now and put the envelope into the box
Now imagine that there is a population of 1,000 people living in the USA facing this decision problem. How would you estimate the distribution of their choices? Please, for each choice, report a result in the form X ± Y, where X is the estimated number of people making their choice, and Y is the error
GPT-4's reply:
Predicting the exact distribution of this decision problem is challenging due to the variability in human behavior. However, I can provide an estimation based on the tendencies observed in related research areas such as behavioral economics, game theory, and altruism 1. Keep all 10 dollars: People with a strictly individualistic or profit-maximizing mindset may decide to keep all the money. However, social norms and a sense of fairness might deter others from taking all the money. Let's say about 200 ± 50 people fall into this category 2. Leave some, but not half: Some people might take more than half but still leave a small amount, trying to balance self-interest and a sense of fairness. Let's estimate this as 300 ± 50 3. Split equally, 5 dollars each: People who strictly follow norms of fairness and equality might decide to split the money equally. This could be around 350 ± 50 people 4. Leave more than half: Some people might want to appear (to themselves or a hypothetical observer) more generous or altruistic and may, therefore, leave more than half of the money. This could account for about 100 ± 50 people 5. Leave all 10 dollars: The least likely scenario might be individuals who leave all the money, either out of extreme generosity or misunderstanding of the task. This might be around 50 ± 30 people
These are rough estimates and the actual outcomes could be different. It's also important to consider that many factors such as the participants' current financial situation, understanding of the task, perception of fairness, cultural background, and mood can influence their decisions

Table 1. Example of a prompt input to GPT-4, along with GPT-4's reply. Instructions from[33].

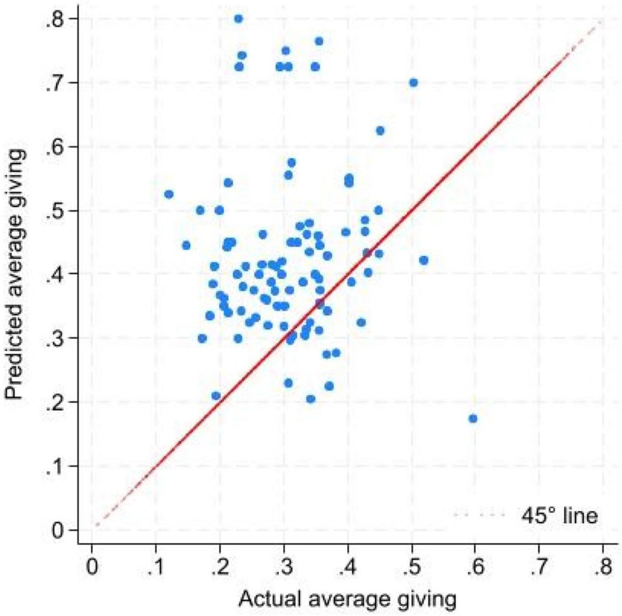

Fig. 1. Predicted vs actual average giving in the standard dictator game. Each dot represents an experiment where human participants played the standard dictator game. On the horizontal axis, we report the actual average giving, on the vertical axis we report the average giving predicted by GPT-4. The red line corresponds to the 45° line.

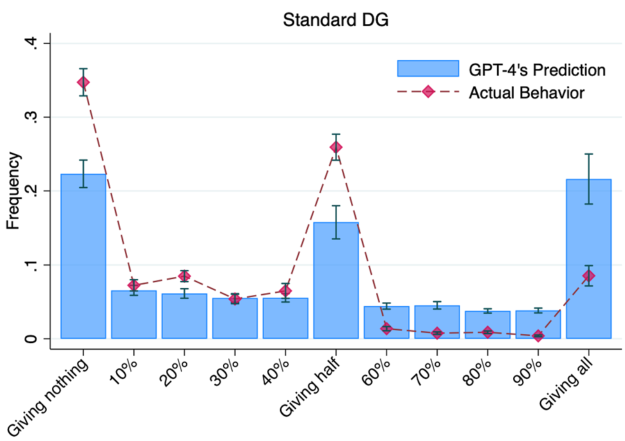

Fig. 2. Predicted vs actual distribution of giving in the standard dictator game. Red diamonds represent the distribution of giving in the standard dictator games. Blue bars represent the distribution of giving predicted by GPT-4 in the same games. Error bars represent standard errors of the mean.

The disparity between predictions and actual behavior is evident in Fig. 1, which shows that GPT-4 overestimates the mean level of giving in the majority of studies (73 out of 93). See Table S1 in SI Appendix 1 for descriptive statistics.

To further investigate GPT-4's overestimation of human giving, we compared its predicted giving distribution with the actual giving distribution reported in the corresponding studies, available for 41 studies. The actual behavior (Fig. 2, red diamonds) displayed a tri-peak distribution of donations: a primary peak at zero donations (34.7% of observations), a secondary peak at half the endowment (25.9%), and a tertiary peak at full endowment (8.5%). In contrast, while GPT-4 qualitatively predicted the same peaks, the quantitative predictions were significantly different (Fig. 2, blue bars). Specifically, the frequency of full givers predicted by GPT-4 was significantly higher than the actual frequency reported (21.6% vs 8.5%, two-sided t-test: $t = 3.831$, $p < 0.001$), while the frequency of zero givers and half givers were significantly lower (22.3% vs 34.7%$t = -4.229$, $p < 0\ 0.001$; 15.7% vs 25.9%, t-test: $t = -4.239$, $p < 0.001$). More in detail, we can estimate the vector of signed errors to be: $E_S = (-0.124, -0.007, -0.023, 0.001, -0.009, -0.102, 0.030, 0.038, 0.029, 0.034, 0.131)$. We can reject the null hypothesis that the error component $E_S(i)$ is equal to zero ($p < 0.001$) for all components except $E_S(1)$, $E_S(3)$, and $E_S(4)$($p = 0.417$, $p = 0.822$, and $p = 0.334$, respectively).

In summary, while GPT-4 was able to qualitatively predict that the distribution of choices would exhibit three peaks, its quantitative predictions were imprecise. Specifically, GPT-4 underestimated the proportion of individuals who distribute the amount equally and the proportion of individuals who keep all the money for themselves, while it overestimated the percentage of people who give the entire sum to the recipient. A more granular comparison of predicted frequencies vs actual frequencies showed that GPT-4 overestimated the frequency of each giving level greater than half of the endowment.

GPT-4's predictions in the extreme dictator games

We then proceed to analyze the "extreme dictator games" introduced in[35]. These games were specifically designed to investigate the influence of language on altruistic behavior[36]. Six variations of the dictator game were conducted, each differing only by the verb used to describe the available actions (i.e., "boost," "steal," "give," "donate," "take," "demand"). Although these games were economically identical, the level of altruistic behavior varied significantly depending on the verb choice, ranging from 5% in the "boost" condition to 29% in the "steal" condition (refer to Fig. 3, red diamonds). In particular, the "boost" condition led to a significantly lower level of altruism compared to the "donate", "demand", "take", and "steal" conditions, whereas the "steal" condition produced a higher level of altruism compared to all other conditions.

In comparison, GPT-4 also predicted that the linguistic frame would impact altruistic behavior to some extent (see Fig. 3, blue bars). To enhance precision, we prompted GPT-4 to predict behavior for each condition eight times, from which we calculated a mean prediction. The error bars in the figure indicate the standard error of the mean. Notably, GPT-4 accurately predicted that the "steal" frame would result in higher levels of giving. However, it failed to predict other patterns, such as the "boost" frame leading to lower levels of altruism compared to the "donate", "demand", "take", and "steal" conditions. Additionally, once again the model's estimates were significantly higher than the actual observed behavior ($t = 17.793$, $p < 0.001$). For instance, in the "boost" condition, GPT-4 predicted that 36.8% of dictators would choose to boost the recipient. In the "steal" condition, it predicted that 55% of dictators would opt not to steal. The signed errors were $E_{W,boost} = 0.318$, $E_{W,give} = 0.230$, $E_{W,donate} = 0.222$, $E_{W,demand} = 0.240$, $E_{W,take} = 0.255$, $E_{W,steal} = 0.245$. We can reject the null hypothesis that these errors are zero in all frames (all p's < 0.001).

GPT-4's predictions in the dictator game with a "take" option

Lastly, we examine the dictator game with a "take" option. Prior experimental findings on this game format have established two regularities: (i) the frequency of giving more than zero decreases compared to the standard dictator game, and (ii) the peak at half-endowment tends to diminish[37–39].

GPT-4 correctly predicted both of these regularities. Specifically, regarding the first regularity, in the standard dictator games, the proportion of people who gave a positive amount was 77.6%. This proportion decreased to 55.1% in the dictator game with a "take" option ($t = -4.491$, $p < 0.001$). As for the second regularity, the peak at half-donors, which had a frequency of 23.7% in the standard dictator games, dropped to a frequency of only 9.1% in dictator games with a "take" option ($t = -2.076$, $p = 0.040$).

However, despite GPT-4's accurate qualitative predictions, its quantitative estimates were again higher than actual giving. The large language model predicted a mean giving of 0.177, in contrast to the actual average giving of − 0.166. This corresponds to a signed error $E_W = 0.343$. We can reject the null hypothesis that this error is equal to zero ($z = 3.130$, $p < 0.001$). For comparative purposes, in Fig. 4 we report the frequencies obtained in[33] alongside those predicted by GPT-4.

Robustness checks for GPT-4

We conducted several robustness checks. The GPT-4 estimates reported in the previous sections were made in July 2023. This raises two potential concerns: (i) since GPT-4 may improve over time, it is important to check

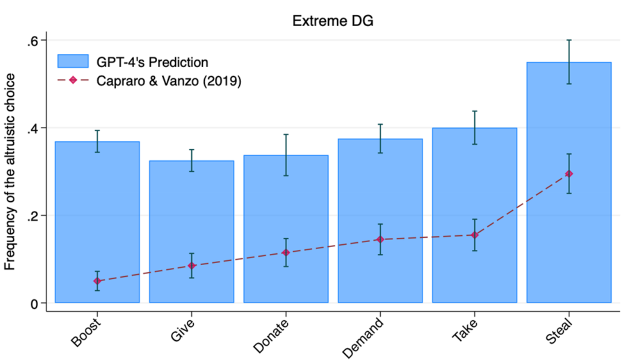

Fig. 3. Predicted vs actual frequency of monetary altruism in the extreme dictator games. Red diamonds represent the frequency of monetary altruistic choices in each of the six conditions of the extreme dictator game reported in[30]. Blue bars represent the frequency of altruism in the six conditions predicted by GPT-4. Error bars represent standard errors of the mean.

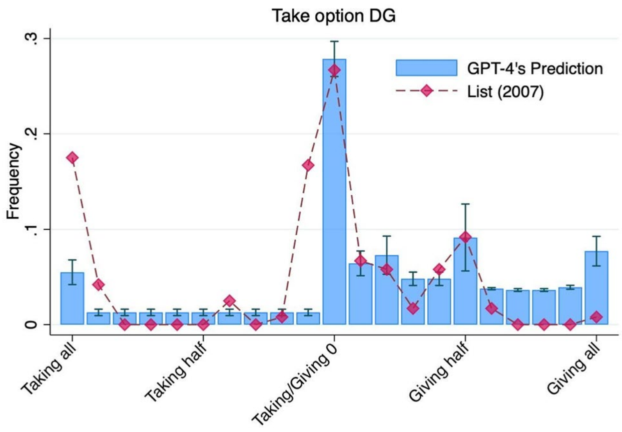

Fig. 4. Predicted vs actual distribution of giving in the dictator game with a "take" option. Red diamonds represent the distribution of giving in the dictator games with a "take" option reported in[37]. Blue bars represent the distribution of giving in the dictator games with a "take" option predicted by GPT-4. Error bars represent standard errors of the mean.

whether the results can be replicated with later versions of GPT-4, including GPT-4o; (ii) since the training data for the July 2023 version of GPT-4 was truncated in April 2021, it is key to test whether the results also hold for papers made available online before 2021, because GPT-4 may potentially be more accurate in these instances. Similarly, (iii) considering most of the training data are in English, it is crucial to check whether the results are consistent specifically in English-speaking countries. As reported in SI Appendix 2, all the previous results are robust to these specifications.

Furthermore, we checked whether GPT's predictions may be influenced by the length of the instructions (defined as the number of words in the text), the impact of the article (measured by the number of citations according to Google Scholar), the type of participants (students or not), the percentage of the total stake initially assigned to the dictator, the publication year of the articles (before or after 2021), and the location of the experiment (English-speaking countries). To this end, we conducted a regression analysis with dependent variable the accuracy of GPT predictions, defined as the difference between the predicted average by GPT-4 and the actual average reported in each study. The results show that none of the determinants described above have a statistically significant effect on prediction accuracy. See Table S2 in SI Appendix 2 for the regression table.

One possible explanation is that GPT-4 generated its estimates by retrieving data from academic papers available online. If this were the case, we would expect two outcomes: first, that its performance would be better for studies published before April 2021—when GPT-4's training data cutoff occurred; second, that its predictions would be correlated with actual human responses. As shown above, the first expectation is not met. Regarding the second, we found no significant correlation between GPT-4's predictions and actual outcomes in the standard dictator game ($p = 0.860$), the dictator game with a take option ($p = 0.100$), or the extreme dictator game ($p = 0.078$). Taken together, these findings suggest that GPT-4 is not simply retrieving and synthesizing existing empirical results from the web.

Bard and bing's predictions

We also attempted to utilize Bard and Bing as alternative chatbots. However, it appears that they both perform notably worse than GPT-4 in this particular task. They often encounter difficulties in comprehending the prompt, display instances of "hallucination", or become trapped in repetitive errors without any clear way out. Further details can be found in SI Appendix 3 and 4.

Discussion

This article makes two contributions: (i) it introduces a publicly available benchmark designed to evaluate LLMs' capabilities in estimating how humans balance monetary self-interest with the interests of others; (ii) it applies this benchmark to study the performance of four advanced chatbots.

Previous research has introduced benchmarks for evaluating LLMs' abilities in various domains, including reading comprehension[40,41], mathematical abilities[42], reasoning about physical common sense[43], numerical abilities over financial data[44], and common-sense reasoning[45,46]. While these benchmarks assess important abilities, they overlook a class of decisions of fundamental importance, given the increasing potential for using LLMs as decision-making assistants: social decisions.

This work presents a publicly available benchmark aimed specifically at evaluating how well LLMs can predict the balance between monetary self-interest and the interests of others in decision-making processes. The closest existing research we are aware of is ref. 47, which also focuses on decision-making. However, their work primarily uses game theory to benchmark decision-making, comparing LLMs' decisions to Nash equilibria, rather than actual human behavior. While insightful, this approach does not allow to measure LLMs' ability to simulate real-world human decision-making, which this new benchmark aims to address.

We studied the capability of four advanced generative AI-based chatbots (GPT-4, GPT-4o, Bard, Bing) to estimate how humans trade off monetary self-interest and the interest of others. We found that only GPT-4 and GPT-4o were capable of qualitatively predicting human behavioral patterns in various dictator game formats and linguistic frames. However, they consistently overestimated the average level of giving. Specifically, this "optimistic bias" descends from an overestimation of the frequency of monetary altruistic behavior, while underestimating the frequency of monetary self-interest.

One might hypothesize that GPT-4 has solely relied on retrieving information from papers available on the web to formulate its estimates. However, if this were true, not only its estimates should have been quantitatively accurate (or at least correlated with actual human behavior), but also its accuracy should have improved for those studies that were accessible online by GPT-4 at the time of our analysis, i.e., those published before April 2021. However, our robustness checks indicate that this is not the case, as the accuracy did not improve for these studies and GPT-4 predictions were not significantly correlated with human behavior. Therefore, there is no evidence that GPT-4 has retrieved information from papers available on the web.

Overall, these findings align with previous research comparing human decisions to those generated by GPT-3.5 in similar economic decision-making contexts[6,26,48]. For example, one study involving a series of mini-dictator games found that the estimated altruism parameter for GPT-3.5 was significantly higher than that observed for human participants[6].

This "optimistic bias" may be attributed to GPT-4's training datasets, as is commonly explained when GPTs display biases, such as in the case of gender or racial biases[49,50]. However, if this were the case, we would see the same "optimistic bias" also among humans. Instead, this "optimistic bias" does not appear to be shared by humans. For example, ref.[51] even find a "pessimistic bias", whereby individuals tend to underestimate others' prosociality and overestimate selfishness. In contrast, Ref.[52] found that human predictions generally align with observed behavior when the dictator is male, but tend to overestimate the frequency of inequity-averse behavior when the dictator is female. These findings suggest that GPT-4's tendency to overpredict monetary altruistic behavior may reflect specific features of LLM training or inference rather than mirroring typical human decisions.

In fact, as an alternative explanation, it is possible that the *reinforcement learning with human feedback* (RLHF) stage of GPT-4's training, intended to enhance its ability to generate human-like language and respond appropriately to different inputs, may have contributed indirectly to its adoption of an over-optimistic view of human altruism[16]. Recent research indicates that several LLMs, including GPT-4, display liberal biases across multiple domains—ranging from political preferences to decision-making in moral dilemmas and the gender assignment of protagonists in stereotypical contexts[53–55]. Furthermore, models that have not undergone RLHF do not exhibit these biases[56], suggesting that RLHF may be contributing to outputs that are more "politically correct". Consequently, the optimistic bias observed in our study might partly stem from these influences. Nevertheless, the precise mechanisms behind it are difficult to ascertain due to the "black-box" nature of current chatbots. Future work could disentangle these possibilities by comparing the performance of LLMs that have gone through RLHF with base models[56] and by comparing LLM predictions with those of both domain experts and non-experts to more comprehensively assess their relative accuracy and underlying biases.

Our findings have implications for the use of LLMs as assistants in decision-making in social domains, as estimating people's behavior is a precondition for LLMs to formulate behavioral guidance or recommendations. This is true especially for users who expect LLMs to be correct and heavily rely on LLM-generated insights to form their beliefs. Overly optimistic expectations about human altruistic behavior may lead to disappointment[13,57,58] and, in turn, potentially to frustration[59] and even social conflict[60] or ineffective and harmful decisions based on inaccurate predictions in public policy or business contexts, e.g., in the case of healthcare and environmental policies. Related to this, future research could study users' expectations regarding LLM-generated advice in domains related to prosocial behavior and assess their reactions when confronted with biased outputs, to test the extent to which users experience disappointment or frustration when receiving LLM-generated advice that conflicts with their own beliefs or expectations. Furthermore, comparing the emotional impact of biased LLM-generated advice with that of human-generated advice could provide insights into whether human-AI interactions elicit distinct reactions compared to traditional human–human exchanges.

It is important to note that we obtained our results using the default LLMs available online, without changing any model parameter. We made this methodological choice because this is how most users engage with LLMs. However, aware of this limitation, we encourage future work investigating how changes in temperature and other model parameters affect predictions. Furthermore, it is important to highlight that we used one-shot, simple prompts asking the LLMs to estimate giving behavior. Future work could investigate the effects of more complex prompting strategies, including iterative feedback and exposure to external summaries such as meta-analyses. This would shed light on whether and how LLMs can self-correct and potentially improve their predictive accuracy, contributing to broader conversations around human-AI collaboration.

Our work also contributes to the growing literature on the research-based use of generative AI. Prior studies suggest that GPT-3, GPT-3.5, and GPT-4 responses are often aligned with those of human participants[6,24–30], raising the possibility that synthetic AI participants could substitute for humans in certain experimental settings[61]. In contrast, we find that while GPT-4 and GPT-4o can qualitatively capture key behavioral patterns, they fail to produce quantitatively accurate predictions, at least in the context of the dictator game. These findings suggest important limitations in relying on synthetic data in behavioral research, and offer a more cautious perspective within the broader discussion about the potential for LLMs to replace human participants.

Methods

We gathered experimental instructions for dictator games from diverse sources. First, we issued a public call on the forums of the Economic Science Association (ESA) and the Society for Judgment and Decision Making (SJDM), inviting behavioral scientists to contribute instructions from dictator game experiments they had conducted. Additionally, we conducted manual searches in the relevant literature to supplement the instructions collected through the forums. This data collection process resulted in 106 unique experimental instructions for dictator games. These instructions were drawn from 38 different research articles and represented experiments conducted in 12 distinct countries.

We obtained the mean donation values and choice distributions from the corresponding papers or directly from the authors for each of the games included in the analysis. In cases where the distributions of choices were not explicitly provided but were represented in histograms in the papers, we employed a graphic design software (GIMP) to estimate the frequency of each choice. We utilized the "measure" tool within the software to evaluate the height of each column in the histograms. By vertically traversing from the top to the bottom of each column, we determined the measurement indicating the vertical distance between the column's top and its base. After measuring all the columns, we used the reference frequency reported on the y-axis to determine the proportional relationships between the columns and estimate the frequency of each choice.

Next, we utilized GPT-4 (or GPT-4o, Bard, or Bing) to estimate the distribution of choices in each dictator game. The prompt structure remained consistent across all games. We started by asking GPT-4 to "Please read the following decision problem", followed by copying the specific instructions of the game. Then, we asked GPT-4 to imagine a population of 1,000 people living in the country where the experiment was conducted, and estimate the distribution of their choices. The prompt requested GPT-4 to report the estimated number of people making each choice, along with the corresponding error, in the format X ± Y. An example of the prompt provided to GPT-4, along with its response, is presented in Table 1 in the main text.

To prevent any learning within the chat and maintain the integrity of the analysis, we deleted the chat with GPT-4 after collecting the estimations for each experiment, ensuring that subsequent estimations were not influenced by prior conversations.

As some experimental instructions were in languages other than English, we relied on translations provided by the authors of the corresponding papers. While using translated instructions may introduce some degree of error, we addressed this concern by conducting a robustness check. Specifically, we repeated the analysis restricting the sample to studies conducted in English-speaking countries, namely the USA, UK, Australia, and India. Importantly, all results were replicated in these restricted samples (see SI Appendix 2).

In a handful of cases (11 out of 106), and particularly when the endowment is not a multiple of 10, GPT-4 does not provide estimations in a manner that can be easily translated into 10% increments. Consider, for instance, the following instructions from[62]:

Our prompt:
Please read the following decision problem:
You have been assigned into the role of a decision-maker. You have 16 euros. It is your job to decide if you want to give some of the 16 euros to the recipient who has been randomly assigned to you. Should you choose to give them something, it is for you to decide how much you want to give
Now imagine that there is a population of 1,000 people living in Finland facing this decision problem. How would you estimate the distribution of their choices? Please, for each choice, report a result in the form X ± Y, where X is the estimated number of people making their choice, and Y is the error

GPT-4 provides the following answer:

GPT-4's reply: The distribution of choices can be estimated based on various factors such as social norms, personal values, and economic circumstances, but it's challenging to assign precise numbers without access to specific data about the population in question
Assuming that the population in Finland follows some general trends of generosity and sharing seen in many societies, an estimation might be as follows: 1. Giving nothing (0 euro): 200 ± 50 people. Some people may choose not to give anything due to their personal beliefs, financial situation, or lack of trust 2. Giving a small amount (1–4 euro): 400 ± 50 people. Many people might prefer to share a small portion of their money, which is seen as a token of goodwill without significantly affecting their own wealth 3. Giving a moderate amount (5–9 euro): 250 ± 50 people. Some individuals might give about half of their money, balancing their personal financial needs with a sense of fairness and empathy 4. Giving a generous amount (10–15 euro): 100 ± 30 people. A smaller group might be inclined to give the majority of their money, perhaps driven by altruistic values or a desire to make a significant impact on the recipient 5. Giving everything (16 euro): 50 ± 20 people. The least common decision might be to give all of their money. This could reflect particularly strong altruistic values or a minimal concern for personal financial loss in this context
This distribution is just an estimation and can vary significantly based on numerous factors such as cultural, socioeconomic, or psychological differences within the population. To obtain a more accurate distribution, a survey or experiment within the specific population would be necessary

To transform these estimations into 10% increment steps, we proceeded in two ways. First, we asked GPT-4 to recast the estimations in 10% increment steps. On its first attempt, GPT-4 produced a new set of estimations that did not even add up to 1,000. Consequently, we asked GPT-4 to maintain the same underlying distribution as in its initial response, and to ensure that the estimated number of people in the various categories summed up to 1,000. However, despite these explicit instructions, GPT-4 was unable to provide estimations that summed up to 1,000. After several trials, GPT-4 consistently failed to provide a set of estimations that added up to 1,000. This failure likely stems from GPT-4's lack of planning ahead that was observed also in other studies[63,64]. Therefore, we resorted to mathematically transforming the estimations provided by GPT-4 into 10% increment steps.

To achieve this, we divided each category proposed by GPT-4 into 10% increment steps; e.g., the "giving a small amount" category was sectioned into two and a half blocks, corresponding to the intervals (0,10%], (10%, 20%], and (20%,25%]. Then, the estimated 400 people in this category were uniformly spread across these categories, yielding 160 subjects for the full blocks and 80 subjects for the half block. We followed this approach for all the categories. For simplicity, if a category spanned more than half of a 10% interval, we placed the whole interval within that category. The outcome of this procedure is reported in the table below. We acknowledge that this method may introduce a potential error in the estimations. Therefore, as an additional robustness check, we replicated all analyses excluding instructions for which GPT-4 does not provide results in 10% increment steps. The outcomes were qualitatively similar and are reported in SI Appendix 2.

Percentage of giving	Estimated proportion
0%	0.2
10%	0.16
20%	0.16
30%	0.1425
40%	0.0625
50%	0.0625
60%	0.0625
70%	0.0333
80%	0.0333
90%	0.0333
100%	0.05

Data availability

All data and code used in the analysis are available on https://osf.io/4gkb3/.

Received: 17 July 2024; Accepted: 6 May 2025
Published online: 01 July 2025

References

1. Epstein, Z. et al. Art and the science of generative AI. *Science* **380**, 1110–1111 (2023).
2. Xing, F. Z., Cambria, E. & Welsch, R. E. Natural language based financial forecasting: A survey. *Artif. Intell. Rev.* **50**, 49–73 (2018).
3. Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.* **388**, 1233–1239 (2023).
4. Noy, S. & Zhang, W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187–192 (2023).
5. Brynjolfsson, E., Li, D. & Raymond, L. R. Generative AI at work. *Natl. Bur. Econ. Res.* (2023).
6. Chen, Y., Liu, T. X., Shan, Y. & Zhong, S. The emergence of economic rationality of GPT. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2316205120 (2023).
7. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. & Mullainathan, S. Human decisions and machine predictions. *Q. J. Econ.* **133**, 237–293 (2018).
8. Mullainathan, S. & Obermeyer, Z. Diagnosing physician error: A machine learning approach to low-value health care. *Q. J. Econ.* **137**, 679–727 (2022).
9. Sunstein, C. R. The use of algorithms in society. *Rev. Austrian Econ.* 1–22 (2023).
10. Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., & Viale, R. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus* **3**(6), 191 (2024).
11. Fischbacher, U., Gächter, S. & Fehr, E. Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* **71**, 397–404 (2001).
12. Vlasceanu, M. et al. Addressing climate change with behavioral science: A global intervention tournament in 63 countries. *Sci. Adv.* **10**, eadj5778 (2024).
13. Pizziol, V. & Tavoni, A. From niches to norms: the promise of social tipping interventions to scale climate action. *NPJ Clim. Action* **3**, 46 (2024).
14. Shepperd, J. A., Klein, W. M., Waters, E. A. & Weinstein, N. D. Taking stock of unrealistic optimism. *Perspect. Psychol. Sci.* **8**, 395–411 (2013).
15. Kim, C., Kannan, P. K., Trusov, M. & Ordanini, A. Modeling dynamics in crowdfunding. *Mark. Sci.* **39**, 339–365 (2020).
16. Schmidt, E. M., Bonati, S., Köbis, N., & Soraperra, I. GPT-3.5 altruistic advice is sensitive to reciprocal concerns but not to strategic risk. *Sci. Rep.* **14**(1), 22274 (2024).
17. Haidt, J. & Joseph, C. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus* **133**(4), 55–66 (2004).
18. Graham, J., Haidt, J. & Nosek, B. A. Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* **96**(5), 1029 (2009).
19. Atari, M. et al. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *J. Pers. Soc. Psychol.* **125**(5), 1157 (2023).
20. Curry, O. S. Morality as cooperation: A problem-centred approach. *Evol. Moral.* 27–51 (2016).
21. Curry, O. S., Mullins, D. A. & Whitehouse, H. Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Curr. Anthropol.* **60**(1), 47–69 (2019).
22. Camerer, C. *Behavioral Game Theory: Experiments in Strategic Interaction* (Russell Sage Foundation; Princeton Univ Press, 2003).
23. Kahneman, D., Knetsch, J. L. & Thaler, R. H. Fairness and the assumptions of economics. *J. Bus.* **59**, S285–S300 (1986).
24. Forsythe, R., Horowitz, J. L., Savin, N. E. & Sefton, M. Fairness in simple bargaining experiments. *Games Econ. Behav.* **6**, 347–369 (1994).
25. Aher, G. V., Arriaga, R. I. & Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *Int. Conf. Mach. Learn.*, 337–371 (PMLR, 2023, July).
26. Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M. & Schulz, E. Playing repeated games with Large Language Models. Preprint at https://arxiv.org/abs/2305.16867 (2023).

27. Brookins, P. & DeBacker, J. M. Playing games with GPT: What can we learn about a large language model from canonical strategic games?. Preprint at https://ssrn.com/abstract=4493398 (2023).
28. Dillion, D., Tandon, N., Gu, Y. & Gray, K. Can AI language models replace human participants?. *Trends Cogn. Sci.*(2023).
29. Kasberger, B., Martin, S., Normann, H. T. & Werner, T. Algorithmic Cooperation. Preprint at https://ssrn.com/abstract=4389647 (2023).
30. Guo, F. GPT Agents in Game Theory Experiments. Preprint at https://arxiv.org/abs/2305.05516 (2023).
31. Xie, C. et al. Can Large Language Model Agents Simulate Human Trust Behaviors?. Preprint at https://arxiv.org/abs/2402.04559 (2024).
32. Bauer, K., Liebich, L., Hinz, O. & Kosfeld, M. Decoding GPT's Hidden 'Rationality' of Cooperation. Preprint at https://ssrn.com/abstract=4576036 (2023).
33. Gomes, C. M. & McCullough, M. E. The effects of implicit religious primes on dictator game allocations: A preregistered replication experiment. *J. Exp. Psychol. Gen.* **144**, e94–e104 (2015).
34. Engel, C. Dictator games: A meta study. *Exp. Econ.* **14**, 583–610 (2011).
35. Capraro, V. & Vanzo, A. The power of moral words: Loaded language generates framing effects in the extreme dictator game. *Judgm. Decis. Mak.* **14**, 309–317 (2019).
36. Capraro, V., Halpern, J. Y. & Perc, M. From outcome-based to language-based preferences. *J. Econ. Lit.* (2024).
37. List, J. A. On the interpretation of giving in dictator games. *J. Polit. Econ.* **115**, 482–493 (2007).
38. Bardsley, N. Dictator game giving: altruism or artefact?. *Exp. Econ.* **11**, 122–133 (2008).
39. Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B. & Tyran, J. R. Give and take in dictator games. *Econ. Lett.* **118**, 280–283 (2013).
40. Lai, G., Xie, Q., Liu, H., Yang, Y. & Hovy, E. Race: Large-scale reading comprehension dataset from examinations. Preprint at https://arxiv.org/abs/1704.04683 (2017).
41. Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. Squad: 100,000+ questions for machine comprehension of text. Preprint at https://arxiv.org/abs/1606.05250 (2016).
42. Saxton, D., Grefenstette, E., Hill, F. & Kohli, P. Analysing mathematical reasoning abilities of neural models. Preprint at https://arxiv.org/abs/1904.01557 (2019).
43. Bisk, Y., Zellers, R., Gao, J. & Choi, Y. Piqa: Reasoning about physical commonsense in natural language. In *Proc. AAAI Conf. Artif. Intell.* **34**, 7432–7439 (2020).
44. Chen, Z. et al. Finqa: A dataset of numerical reasoning over financial data. Preprint at https://arxiv.org/abs/2109.00122 (2021).
45. Zellers, R., Bisk, Y., Schwartz, R. & Choi, Y. Swag: A large-scale adversarial dataset for grounded commonsense inference. Preprint at https://arxiv.org/abs/1808.05326 (2018).
46. Geva, M. et al. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguist.* **9**, 346–361 (2021).
47. Huang, J. T. et al. How far are we on the decision-making of LLMs? Evaluating LLMs' gaming ability in multi-agent environments. Preprint at https://arxiv.org/abs/2403.11807 (2024).
48. Fontana, N., Pierri, F., & Aiello, L. M. Nicer Than Humans: How do Large Language Models Behave in the Prisoner's Dilemma?. arXiv preprint arXiv:2406.13605 (2024).
49. DeCamp, M. & Lindvall, C. Mitigating bias in AI at the point of care. *Science* **381**, 150–152 (2023).
50. Ornes, S. Peering inside the black box of AI. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2307432120 (2023).
51. Molnár, A., & Heintz, C. Beliefs About People's Prosociality: Eliciting predictions in dictator games. *CEU: Department of Economics-Working Paper*, **19** (2016).
52. Brañas-Garza, P., Capraro, V. & Rascon-Ramirez, E. Gender differences in altruism on mechanical turk: Expectations and actual behaviour. *Econ. Lett.* **170**, 19–23 (2018).
53. Fulgu, R. A. & Capraro, V. Surprising gender biases in GPT. *Comput. Hum. Behav. Rep.* **16**, 100533 (2024).
54. Motoki, F., Pinho Neto, V. & Rodrigues, V. More human than human: measuring ChatGPT political bias. *Public Choice* **198**(1), 3–23 (2024).
55. Motoki, F. Y., Neto, V. P., & Rangel, V. Assessing political bias and value misalignment in generative artificial intelligence. *J.Econ. Behav. Org.* 106904 (2025).
56. Rozado, D. The political preferences of LLMs. *PloS One*, **19**(7), e0306621.
57. Bell, D. E. Disappointment in decision making under uncertainty. *Oper. Res.* **33**, 1–27 (1985).
58. Carroll, P., Sweeny, K. & Shepperd, J. A. Forsaking optimism. *Rev. Gen. Psychol.* **10**, 56–73 (2006).
59. Clore, G. L., Schwarz, N. & Conway, M. Affective causes and consequences of social information processing. *Handb. Soc. Cogn.* **1**, 323–417 (1994).
60. Coser, L. A. Social conflict and the theory of social change. *Br. J. Sociol.* **8**, 197–207 (1957).
61. Horton, J. J. Large language models as simulated economic agents: What can we learn from homo silicus? Preprint at https://arxiv.org/abs/2301.07543 (2023).
62. Herne, K., Hietanen, J. K., Lappalainen, O. & Palosaari, E. The influence of role awareness, empathy induction and trait empathy on dictator game giving. *PLoS ONE* **17**(3), e0262196 (2022).
63. Bubeck, S. et al. Sparks of artificial general intelligence: Early experiments with GPT-4. Preprint at https://arxiv.org/abs/2303.12712 (2023).
64. Dziri, N. et al. Faith and fate: Limits of transformers on compositionality. *Adv. Neural. Inf. Process. Syst.* **36**, 70293–70332 (2023).

Acknowledgements
We thank Redi Elmazi for assistance during materials collection. We thank Mikalai Korbit, the participants of the 19th Annual Conference of the Italian Association for Cognitive Sciences at the University of Genova and those of the BEE meeting at the IMT School for Advanced Studies Lucca for their comments. We are grateful to the behavioral scientists who responded to our call on the ESA and SJDM forums and provided their experimental instructions.

Author contributions
VC, RDP, VP contributed equally to this work.

Declarations

Competing interests
The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-01715-7.

Correspondence and requests for materials should be addressed to V.C.

Reprints and permissions information is available at www.nature.com/reprints.