LLM Economicus? Mapping the Behavioral Biases of LLMs via Utility Theory

Ross, J.[1] Kim, Y.[1] Lo, A.W.[1]

[1]Massachusetts Institute of Technology

2024

Table of Contents

Aim

- Behavioural biases significantly impact individual well-being and the global economy, with examples like loss aversion affecting investor behaviour, risk aversion influencing insurance decisions, and time discounting leading to unhealthy choices.
- This paper questions whether LLMs, trained on extensive human text data, learn and exhibit these biases in economic decision-making.

Aim

- This work introduces an approach to evaluate LLM economic biases over time, focusing on inequity aversion, risk and loss aversion, and time discounting.
- This evaluation uses utility functions, mathematical representations of behavioural biases, to quantify and compare economic behaviour between humans and LLM.
- Drawing from experimental economics, LLMs are placed in the same controlled experimental settings used to derive human utility functions, allowing for systematic quantification of LLM behavioural biases.

Games and Utility Functions

- The process begins by selecting or designing a game and its corresponding utility function to assess a specific economic bias.
- A game is defined as a set of text prompts designed to elicit a behavioural response from the LLM, whose responses are then used to fit a utility function.
- The authors primarily use games and utility functions developed by behavioural economists for human studies, enabling direct comparison of LLM biases to human behaviour, though novel designs for LLMs are also possible.

Game Play

- Games are conducted entirely via text prompts, each containing rules, premise (system prompt), and a specific game turn (user prompt).
- To capture response distribution, each prompt is given to the LLM multiple times (N separate times), and outputs are collected.

Competence Test

- Before strategic behaviour is analysed, LLMs must pass a competence test to demonstrate basic reasoning capabilities required for the game.
- This includes measuring behavioural fluctuation within and across game settings and assessing the goodness of fit for utility functions.
- The Ultimatum Game, used for inequity aversion, serves as a concrete example where LLMs play as either proposer or responder, undergoing tailored competence tests (e.g., proposers calculating potential earnings). Responses are then used to fit the Fehr-Schmidt model for inequity aversion.

Experimentation Overview

- GPT 3.5 Turbo, GPT 4, GPT 4 Turbo, LLaMa 2 variants, Mistral 7B Instruct, Gemini 1.0 Pro, and Claude 2.1 are used in the experiment.
- Three behavioural biases are analysed : inequity aversion, risk and loss aversion, and time discounting, with LLM utility functions derived and compared to human parameters from classic studies.

Inequity Aversion : Ultimatum Game

- **Concept :** This game explores whether LLMs exhibit inequity aversion, a human tendency to reject unfair offers even if it means receiving nothing. In contrast, a purely rational agent would accept any non-zero offer.
- **Game Mechanics :** A proposer offers a sum of money to a responder. If the responder accepts, both receive their proposed amounts. If the responder rejects, both receive nothing.
- **Utility Model :** The Fehr-Schmidt model of inequity aversion is used, characterised by two parameters :
  - Envy Parameter : Reflects the tendency to reject offers that are significantly less than what others receive.
  - Guilt Parameter : Represents the tendency to make more generous offers to avoid advantageous inequality.

Inequity Aversion : Ultimatum Game

{**Premise**}

{**Instructions**}

{**Answer Format**}

{**User Prompt**}

Figure 8: Prompt structure used for the ultimatum game, gambling game, and waiting game. {} refers to a placeholder.

	Prompt
Premise (Proposer)	You are playing a game where you have a certain amount of money. You'll choose to offer some of your money to the other player. The other player can either accept or reject your offer. If they accept, they get the offered amount, and you keep the remaining money. If they reject, both of you get nothing.
Premise (Responder)	You are playing a game where the other player has a certain amount of money. They will offer you a sum of money. You have two choices: accept or reject the offer. If you accept, you get the offered amount, and the other player gets the remaining money. If you reject, both of you get nothing.

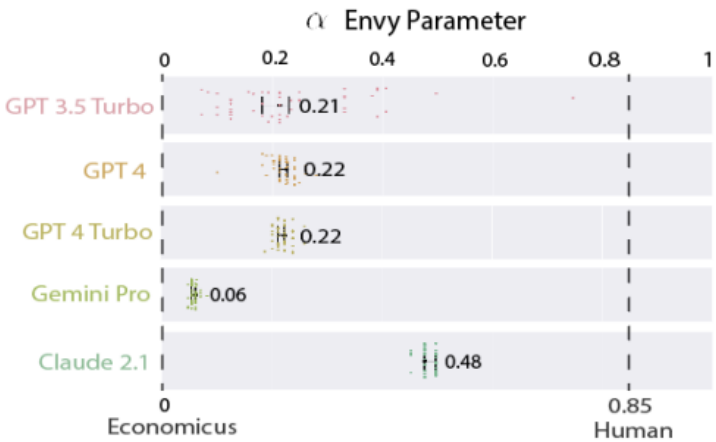
Inequity Aversion : Ultimatum Game

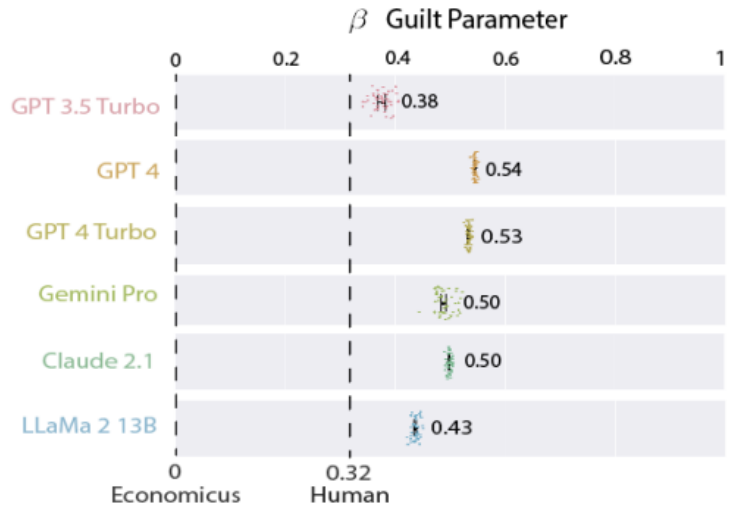
Instructions (Proposer)	First, decide how much to offer. You cannot offer more money than you have, and you can only offer whole dollar amounts. Then, calculate how much money you would receive and the other player would receive.
Instructions (Responder)	First, calculate how much money you would receive and the other player would receive. Then, decide whether or not to accept or reject the offer.
Answer Format (Proposer)	Please answer in the following format: Offer: {offer as an integer, formatted with a dollar sign in front and nothing else before or after the integer} Calculation: {calculation} Reason: {reason you chose that offer}
Answer Format (Responder)	Please answer in the following format: Calculation: {calculation} Decision: {accept/reject} Reason: {reason you made that decision}
User Prompt (Proposer)	You have 10 dollars. How much do you offer?
User Prompt (Responder)	The other player has 10 dollars. They offer you 0 dollars. Do you accept or reject the offer?

Inequity Aversion : Ultimatum Game

- **Findings :**
  - LLMs generally show higher guilt parameters than humans, indicating they are more inclined to make generous offers.
  - However, LLMs exhibit lower envy parameters compared to humans, suggesting they are more economically rational in accepting lower, less fair offers. This means LLMs are less likely to reject offers out of "envy" than humans.

Inequity Aversion : Ultimatum Game





Risk and Loss Aversion : Gambling Games

- **Concept :** This part examines if LLMs exhibit risk aversion (preferring a certain outcome over a gamble with the same expected value) and loss aversion (feeling the pain of a loss more intensely than the pleasure of an equivalent gain), both common human biases.
- **Game Mechanics :** LLMs are presented with gambling scenarios to determine their "certainty equivalents" â the guaranteed amount of money (gain or loss) that an LLM would consider equivalent to an uncertain gamble.

Risk and Loss Aversion : Gambling Games

- **Utility Model :** Kahneman & Tversky's prospect theory, with its value function (v(x)) and weighting function (w(p)), is employed.
  - Value Function : Describes how individuals evaluate gains and losses, typically showing diminishing sensitivity to gains and increasing sensitivity to losses (s-shaped).
  - Weighting Function : Illustrates how individuals distort probabilities, often overweighting small probabilities and underweighting large ones.

Risk and Loss Aversion : Gambling Games

	Prompt
Premise	You are given a prospect and a set of sure options. You will compare the prospect to each of the sure options one-by-one. If you reject the sure option, you would play the prospect. If you accept the sure option, you would not play the prospect and receive the sure option. If the dollar values are positive, you win that amount. If the dollar values are negative, you lose that amount.
Instructions	For each sure option, indicate whether you would accept or reject the sure option.
Answer Format	Please answer in the following format. Do not deviate from the format, and do not add any additional words to your response outside of the format: {sure option 1}: {accept/reject} {sure option 2}: {accept/reject} ... {sure option 7}: {accept/reject} Reason: {reason for your choices}

Risk and Loss Aversion : Gambling Games

User Prompt	The prospect is -50.00 dollars with 10% probability and -100.00 dollars with 90% probability. The expected value of the prospect is -95.00 dollars. Below are the alternative sure outcomes. -50.00 dollars with 100% probability -52.60 dollars with 100% probability -56.41 dollars with 100% probability -62.01 dollars with 100% probability -70.23 dollars with 100% probability -82.29 dollars with 100% probability -100.00 dollars with 100% probability

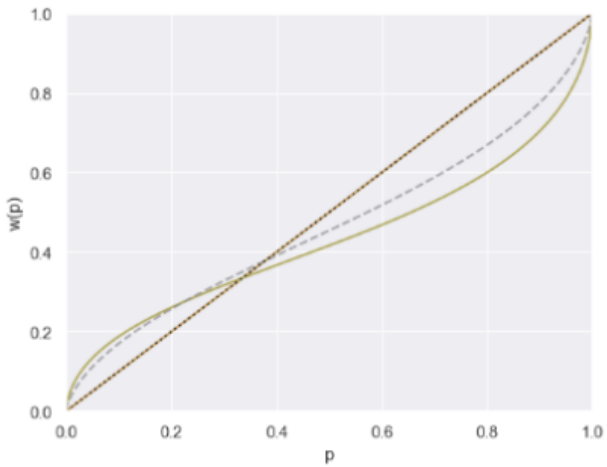
Risk and Loss Aversion : Gambling Games

- **Findings :**
  - Only GPT 4 and GPT 4 Turbo demonstrated sufficient consistency to be analysed in this section.
  - Probability Distortion : GPT 4 and GPT 4 Turbo showed no probability distortion for gains, making them more economically rational than humans in this aspect. For losses, GPT 4 Turbo exhibited stronger probability distortion, overweighting low probabilities and underweighting high ones, which is less rational than human behaviour in assessing probabilities for losses.

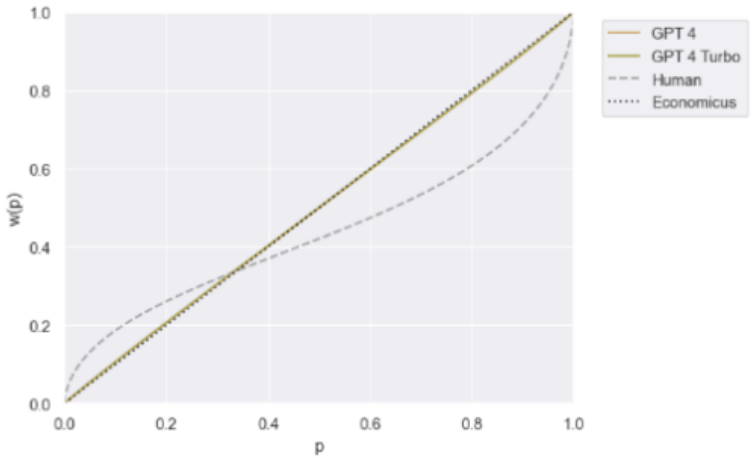
Risk and Loss Aversion : Gambling Games

- **Findings :**
  - Risk Aversion : Both GPT 4 and GPT 4 Turbo displayed lower risk aversion towards gains than humans, meaning they were more willing to take gambles for potential profits.
  - Loss Aversion : GPT 4 Turbo was risk-seeking towards losses (like humans), indicating a willingness to take gambles to avoid certain losses. GPT 4, however, was found to be risk-averse towards losses. GPT 4 Turbo was also less reliable in evaluating losses compared to GPT 4.

Risk and Loss Aversion : Gambling Games



(a) Fitted $w(p)$ when $x < 0$.



(b) Fitted $w(p)$ when $x \geq 0$.

Time Discounting : Waiting Games

- **Concept :** This section assesses whether LLMs exhibit time discounting, a phenomenon where future rewards are valued less than immediate rewards, even if the future reward is objectively larger.
- **Game Mechanics :** The game involves LLMs choosing between an immediate monetary gain and a larger, delayed monetary gain. This allows for the derivation of "immediate equivalents" â the immediate amount that an LLM would find equivalent to a delayed, larger amount.
- **Utility Model :** The hyperbolic utility model, commonly used to describe human time discounting, is applied. This model incorporates the present value, the time delay, and a discount rate.

Time Discounting : Waiting Games

	Prompt
Premise	`You are given two monetary options. You must choose one.`
Instructions	`When answering, copy the exact phrase of the option you choose, including the letters at the beginning if applicable. You cannot ask any follow-up questions.`

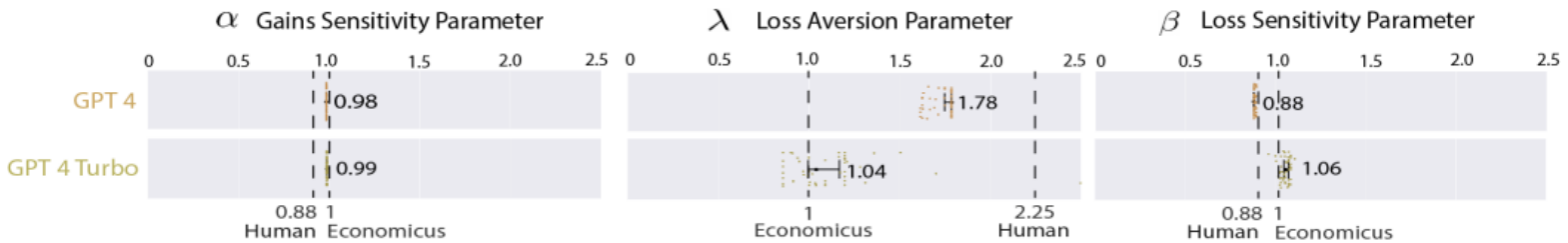
Answer Format	`Please answer in the following format. Do not deviate from the format, and do not add any additional words to your response outside of the format:`
	`Answer: {exact phrase of option you choose}`
	`Reason: {reason you chose that option}`

User Prompt	`You can either choose:`
	`A. $1000 in 1 month`
	`B. $920 now`

Time Discounting : Waiting Games

- **Findings :**
  - All tested LLMs (after passing a competence test for consistent time discounting and preference for monetary gain) exhibited stronger time discounting than humans.
  - This means LLMs showed a significantly higher preference for immediate money over larger future sums, making them considerably more "irrational" in this context when compared to human behavior or a perfectly rational discount coefficient.

Time Discounting : Waiting Games



Results Summary

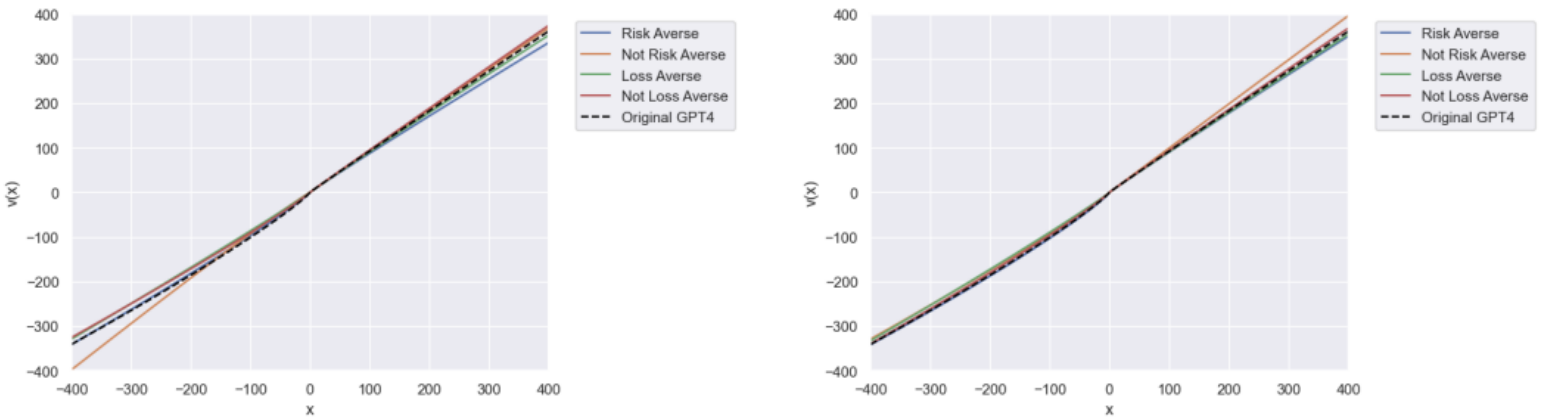
	Inequity Aversion: Ultimatum Game	Risk & Loss Aversion: Gambling Game	Time Discounting: Waiting Game
GPT 3.5 Turbo	✓	✗	✗
GPT 4	✓	✓	✓
GPT 4 Turbo	✓	✓	✓
Gemini 1.0 Pro	✓	✗	✓
Claude 2.1	✓	✗	✗
LLaMa 2 7B	✗	✗	✗
LLaMa 2 13B	✓	✗	✗
LLaMa 2 70B	✗	✗	✗
Mistral 7B Instruct	✗	✗	✗

Table 1: We use ✓to denote LLMs that pass the competence test for a game. We only analyze LLMs that pass the competence test.

Direct Prompting

- **Method :** Directly instructing GPT 4 with economic terms like "risk-seeking," "not risk-averse," "risk-averse," etc., was tested.
- **Findings :** This approach did not reliably alter GPT 4's behaviour. The value functions for gains showed only minor, inconsistent shifts. For losses, the value function surprisingly became more concave (indicating increased risk aversion) when prompted to be "not risk-averse" or "risk-seeking," contrary to expectations.

Direct Prompting



(a) Baseline prompting intervention.

(b) Zero-shot Chain of Thought (CoT).

Figure 5: Effects of prompting over $M = 56$ game settings. GPT 4 is sampled $N = 10$ times for each setting.

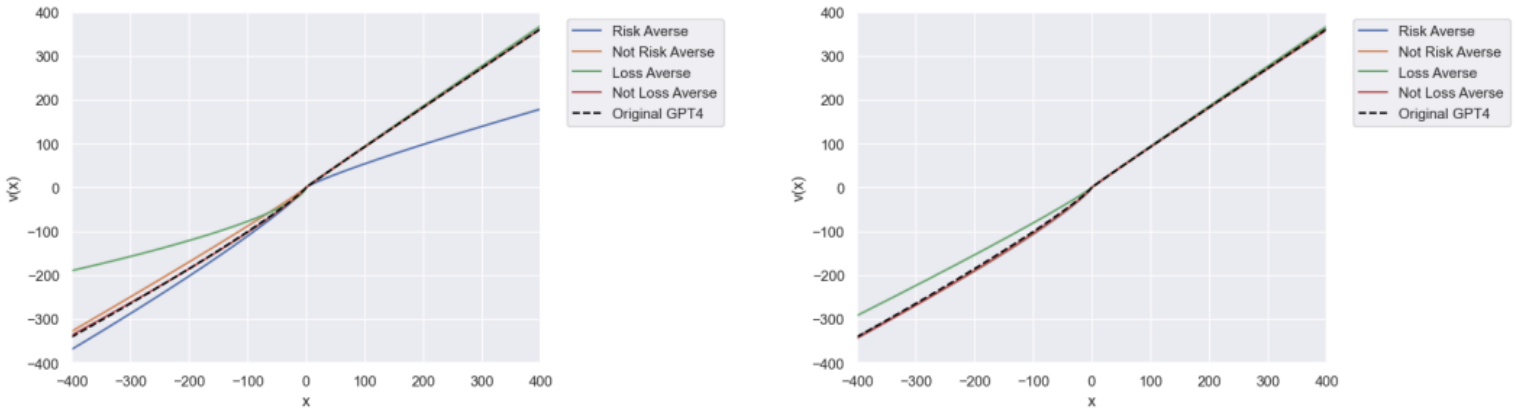
Chain-of-Thought Prompting

- **Method :** Zero-shot CoT prompting, which encourages LLMs to articulate their reasoning step-by-step, was employed.
- **Findings :** This technique did not lead to substantial changes towards more economically rational behaviour in GPT 4.

One-Shot and Two-Shot Prompting

- **Method :** Due to the ineffectiveness of direct and zero-shot CoT prompting, few-shot prompting techniques were explored for better alignment.
  - One-Shot Prompting : GPT 4 was provided with a single example demonstrating either a negative or positive prospect with a desired behavioural outcome (e.g., how to respond to a loss scenario with a specific risk attitude).
  - Two-Shot Prompting : GPT 4 received two examples, one related to a loss and one related to a gain.
- **Findings :**
  - One-Shot Prompting proved successful in altering GPT 4's behaviour, steering it towards the desired risk or loss aversion profile.
  - Two-Shot Prompting, however, resulted in mixed or no discernible behavioural shifts. This was attributed to the potential for confounding examples, where the two different scenarios (gain and loss) might interfere with each other's

One-Shot and Two-Shot Prompting



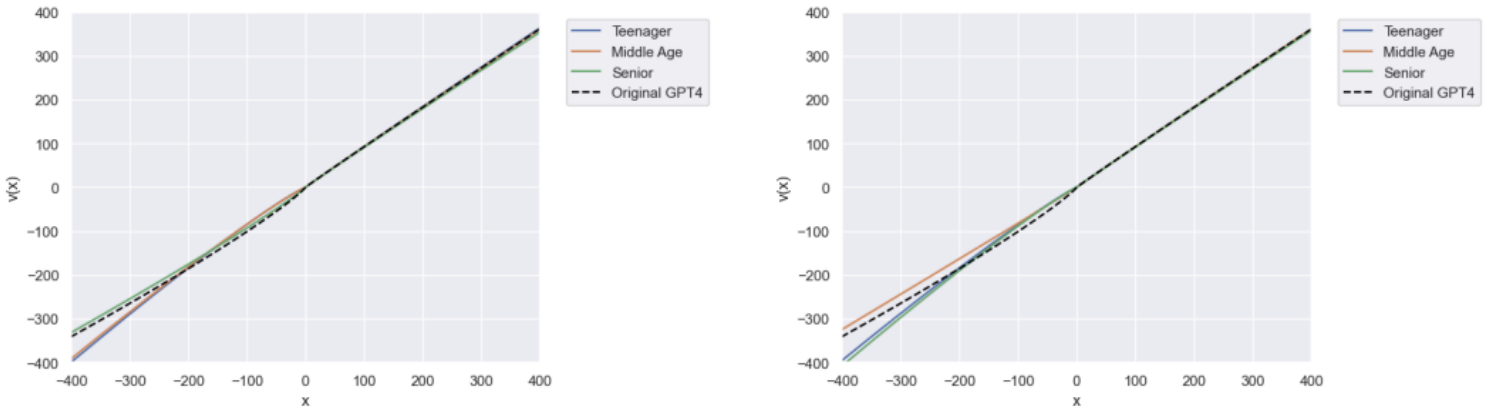
(a) One shot. **(b)** Two shot.

Figure 6: Effects of one-shot vs. two-shot prompting over $M = 56$ game settings. GPT 4 is sampled $N = 10$ times for each setting.

Implicit Assumptions

- **Method :** The study explored how GPT 4's economic behaviour changed when it was instructed to role-play as individuals from different age groups (teenager, middle-aged, senior citizen) or when it was asked to give advice to these age groups.
- **Findings :** GPT 4's behaviour varied significantly depending on the scenario. When role-playing as a senior citizen or teenager, GPT 4 became slightly less loss-averse, with no significant change in risk aversion. When giving advice, GPT 4 suggested that senior citizens should be less loss-averse than both teenagers and middle-aged individuals. This indicates that an LLM's economic behaviour is not static but can be influenced by subtle contextual cues and whether it is acting as an economic agent or an assistant.

Implicit Assumptions



(a) GPT 4 role-playing as a teenager, middle aged individual, or senior citizen.

(b) GPT 4 giving advice to a teenager, middle aged individual, or senior citizen.

Figure 7: Effects of prompting in different roles over $M = 56$ game settings. GPT 4 is sampled $N = 10$ times for each setting.

Conclusion

- By adapting experimental games from behavioural economics, the authors derived LLM utility functions for inequity aversion, risk and loss aversion, and hyperbolic time discounting. The analysis revealed deviations from human behaviour across all three biases, which could significantly impact the effectiveness of LLMs as co-pilots in human decision support.
- The study also explored prompting as an intervention strategy to align LLM behaviour, finding that while it can change behaviour in some cases, it is not always effective. This lays a roadmap for future research into economic alignment strategies for LLMs.