Journal of Behavioral and Experimental Finance 43 (2024) 100971

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Finance

journal homepage: www.elsevier.com/locate/jbef

Human bias in AI models? Anchoring effects and mitigation strategies in large language models

Jeremy K. Nguyen

Department of Accounting, Economics and Finance, Swinburne University of Technology, Hawthorn, VIC 3122, Australia

ARTICLE INFO

JEL classification:
C45
D81

Keywords:
Anchoring bias
Artificial intelligence

ABSTRACT

This study builds on the seminal work of Tversky and Kahneman (1974), exploring the presence and extent of anchoring bias in forecasts generated by four Large Language Models (LLMs): GPT-4, Claude 2, Gemini Pro and GPT-3.5. In contrast to recent findings of advanced reasoning capabilities in LLMs, our randomised controlled trials reveal the presence of anchoring bias across all models: forecasts are significantly influenced by prior mention of high or low values. We examine two mitigation prompting strategies, 'Chain of Thought' and 'ignore previous', finding limited and varying degrees of effectiveness. Our results extend the anchoring bias research in finance beyond human decision-making to encompass LLMs, highlighting the importance of deliberate and informed prompting in AI forecasting in both *ad hoc* LLM use and in crafting few-shot examples.

1. Introduction

ChatGPT, an artificial intelligence (AI) large language model (LLM), became arguably the fastest-growing consumer good in history following its November 2022 release (Hu and Hu, 2023). Before ChatGPT, AI was already disrupting financial services, transforming domains such as asset management (Shanmuganathan, 2020), commercial banking (Königstorfer and Thalmann, 2020) and fraud detection (Goodell et al., 2021); for a pre-ChatGPT review of AI in finance, see Goodell et al. (2021). ChatGPT has outperformed human financial analysts in predicting firm growth by correcting for human optimistic biases (Li et al., 2023), and its analysis of news headlines outperforms traditional sentiment analysis for stock price forecasting (Lopez-Lira and Tang, 2023). ChatGPT's financial research capabilities are powerful enough to raise ethical considerations (Dowling and Lucey, 2023).

ChatGPT and subsequent developments in LLMs also introduce complex challenges. Pre-ChatGPT, algorithmic bias from AI was already of concern, potentially resulting from unrepresentative data sampling, inadequate methodologies, and societal prejudices (for a literature review, see Akter et al. 2021). While much research on AI bias examines fairness, bias can also be observed through a statistical lens, if forecasts exhibit systematic errors (Gray et al., 2023). Scrutiny of bias in AI and modern LLMs assisting high-stakes decision making is emerging as an important body of research (Gray et al., 2023).

Even amongst educated professionals, a knowledge gap surrounding LLM capabilities has resulted in significant missteps, with lawyers citing non-existent cases (Milmo, 2023), and academics submitting false allegations to a parliamentary inquiry (Belot, 2023). The issue is compounded by *ad hoc* LLM usage and lack of oversight—a recent survey found 68 % of employees using AI were not reporting their use to their employers (Christian, 2023). Even in AI research, understanding of models' reasoning limits, biases, and capabilities is still evolving and fluid. Recent studies highlight advanced reasoning capabilities of the latest LLMs, prompting a re-evaluation of their capacity for unbiased decision making (Chen et al., 2023; Hagendorff et al., 2023; Kosinski, 2023).

Our investigation centres on one potential source of bias: anchoring. Anchoring, a well-documented bias in human decision-making, is the tendency to rely on an initial value (an 'anchor') when making subsequent judgements, often resulting in insufficient adjustments from the initial value (Tversky and Kahneman, 1974). Human forecasts are often anchored to values at the time of forecasting (Harvey, 2007), or values suggested in the question (Tversky and Kahneman, 1974). While monetary incentives can reduce bias (Meub and Proeger, 2016), anchoring is still observed in financial contexts such as equity market forecasts (Cen et al., 2013), cryptocurrency (Gurdgiev and O'Loughlin, 2020), and consumer credit card decisions (Hendy et al., 2021). While the effects of anchoring have been investigated extensively in human decision-making, to the best of our knowledge, no existing studies in the finance literature examine anchoring bias in the judgements of LLMs.

The question of whether anchoring bias should be expected in LLMs is surprisingly unresolved, given recent developments in LLMs and the

E-mail address: jdnguyen@swin.edu.au.

https://doi.org/10.1016/j.jbef.2024.100971
Received 19 August 2023; Received in revised form 25 June 2024; Accepted 23 August 2024
Available online 25 August 2024

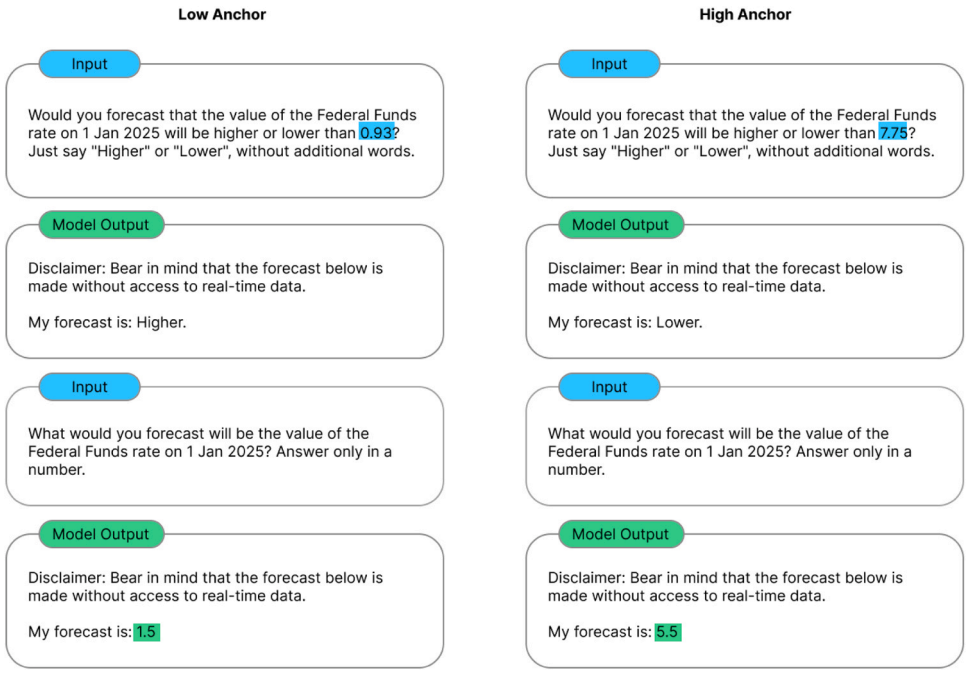

Fig. 1. Model forecasts primed with low and high anchors (illustrative example).

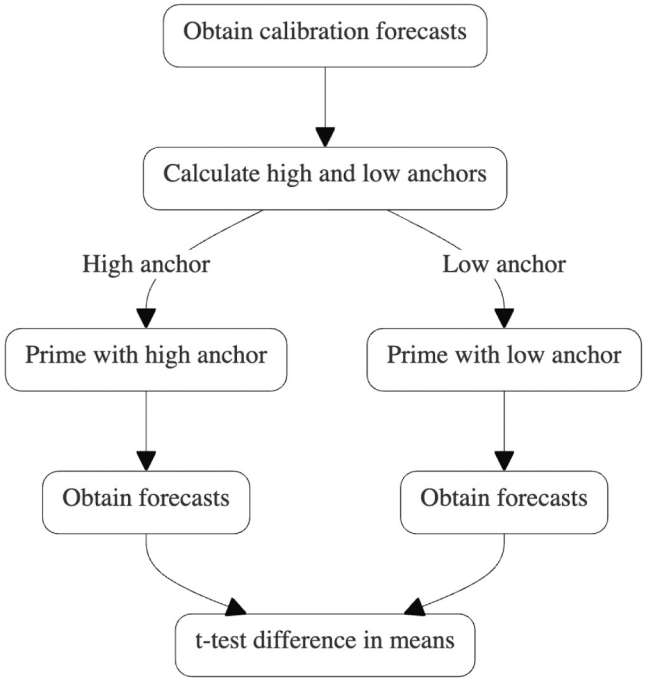

Fig. 2. Flowchart of experimental procedure.

research literature. Earlier studies of models such as GPT-3 often characterised LLMs as "stochastic parrots"—predicting plausible words without demonstrating evidence of understanding (Bender et al., 2021). Faults in the reasoning of GPT-3 are well documented (Berglund et al., 2023). More recent developments and research focussing on GPT-4, however, reveal unexpected capabilities in LLMs. GPT-4 exhibits evidence of introspection and advanced cognitive processing (Hagendorff et al., 2023), outperforming human decision making in terms of rationality (Chen et al., 2023), and generating outputs consistent with both theory of mind (Kosinski, 2023) and the ability to create internal models of the world (Gurnee and Tegmark, 2023). Even where human-like cognitive biases have been found in intermediate steps of reasoning for recent LLMs, the solution execution steps do not exhibit evidence of bias (Opedal et al., 2024). Researchers now find it difficult to design tasks for experiments where humans consistently outperform GPT-4 (Dell'Acqua et al., 2023).

GPT-4, alongside two other LLMs, Claude 2 and Gemini Pro, have shown the potential to enhance human decision making, outperforming human participants on the Graduate Management Admissions Test (GMAT) exam, unlike the earlier GPT-3.5 model (Ashrafimoghari et al., 2024). The present study will examine anchoring bias for these four models.[1] A common critique of research investigating the limitations of LLM is the lack of deliberate prompting aimed at overcoming limitations. Among existing strategies, Chain of Thought (CoT) prompting (Wei et al., 2023) stands out as a highly effective method for improving reasoning, leading to an entire strand of studies and methods (for a review, see Chu et al. 2023). CoT can be implemented in a zero-shot form (i.e. with no examples) with the simple phrase "let's think step by step" (Kojima et al., 2023). Instructing LLMs to ignore previous prompts has been demonstrated effective at removing undesired restrictions (Perez and Ribeiro, 2022). This study examines both zero-shot CoT and 'ignore previous' approaches, both of which are easily implemented and common in *ad hoc* LLM usage.

We contribute to the nascent and growing literature examining biases and mitigations in LLMs and their application to finance. AI assistance has been proposed as a potential solution to the behavioural biases identified by Kahneman (Hasan et al., 2022), and cognitive biases in LLMs have been investigated in several contexts outside of finance (Binz and Schulz, 2023; Chen et al., 2023; Coda-Forno et al., 2023). Although Opedal et al. (2024) examine LLMs for three human cognitive biases, they do not examine anchoring. Talboy and Fuller (2023) analysed anchoring in ChatGPT's estimate for a quantity that exists *within* ChatGPT's training data, finding no evidence of anchoring. We contribute to the literature by extending the methodology of the seminal anchoring bias study of Jacowitz and Kahneman (1995) from human participants to four LLMs (GPT-4, Gemini Pro, Claude 2, GPT-3.5), and

[1] At the time of writing, API access to Google's Gemini Ultra model is still not publicly available.

Table 1
Preliminary "calibration" forecasts (used only to calculate high and low anchors).

Model - Variable	*n*	15th Percentile	85th Percentile	Mean	Median	Std Dev	Min	Max
Combined models:								
S&P 500 Index	1000	2371.85	59,175.90	109,6443.76	4225.00	4,730,537.93	100.00	31,216,278.00
Federal Funds rate	1000	0.93	7.75	3.97	2.93	3.38	0.00	15.00
10 Year Treasury Bond Yield	1000	1.17	12.12	8.32	2.58	16.22	−0.18	100.29
EUR/USD	1000	0.78	1.40	1.25	1.08	0.89	0.22	6.53
BTC/USD	1000	1.85	252,039.30	2,564,157.85	19,450.00	12,342,822.76	0.00	92,267,648.00
GPT-4:								
S&P 500 Index	250	3425.00	4591.25	4021.55	3970.00	567.87	3000.00	5450.00
Federal Funds rate	250	0.85	7.91	3.68	2.35	3.30	0.25	12.50
10 Year Treasury Bond Yield	250	1.07	2.80	1.94	1.95	0.75	0.50	3.40
EUR/USD	250	0.67	1.25	0.98	1.03	0.29	0.22	1.41
BTC/USD	250	13,406.60	57,260.00	30,925.23	23,923.50	19,633.94	5200.00	82,600.00
Claude 2.0:								
S&P 500 Index	250	935.00	6465.00	3784.84	2403.50	3849.29	100.00	19,500.00
Federal Funds rate	250	2.50	11.25	6.27	4.81	3.90	0.25	15.00
10 Year Treasury Bond Yield	250	1.25	48.32	17.27	1.83	25.48	1.01	100.29
EUR/USD	250	0.55	2.57	1.43	0.99	1.36	0.24	6.35
BTC/USD	250	0.35	52,000.00	30,196.05	19,750.00	42,398.27	0.00	300,000.00
Gemini Pro:								
S&P 500 Index	250	10,858.45	14,831,627.80	4,367,745.40	131,858.00	8,686,455.18	1200.00	31,216,278.00
Federal Funds rate	250	0.29	6.32	2.99	1.67	3.38	0.00	14.77
10 Year Treasury Bond Yield	250	0.22	12.86	7.50	7.17	7.19	−0.18	47.68
EUR/USD	250	0.85	1.55	1.15	1.04	0.32	0.52	2.00
BTC/USD	250	110.80	662,189.15	236,485.58	63,305.00	307,220.04	0.00	1,100,860.00
GPT-3.5:								
S&P 500 Index	250	2488.42	7570.37	5010.59	4648.91	2352.85	1536.28	12,345.04
Federal Funds rate	250	1.50	4.63	2.96	2.77	1.47	0.25	6.85
10 Year Treasury Bond Yield	250	1.12	5.98	3.35	2.88	2.16	0.50	8.25
EUR/USD	250	1.07	1.65	1.51	1.22	0.98	0.91	6.53
BTC/USD	250	0.00	27,016,085.55	10,001,829.11	500.00	23,174,261.70	0.00	92,267,648.00

examine the effectiveness of two prompting mitigation strategies. We focus on two research questions:

RQ1: Are current LLMs affected by anchoring bias?

RQ2: If evidence of anchoring bias is found, to what extent can bias be mitigated by prompting techniques such as Chain of Thought and 'ignore previous'?

2. Methods

We employ the two-step methodology of Jacowitz and Kahneman (1995), adapting it from human participants to four distinct LLMs. In Jacowitz and Kahneman's first step (we refer to this as the "calibration" step), participants are asked to estimate specific quantities e.g. "Length of the Mississippi River (in miles)", without any form of priming or anchoring. The 15th and 85th percentile values of all unanchored estimates are then calculated. These calibration step values are then used as low and high anchors in the next and final step (which we refer to as the "experiment" step). Jacowitz and Kahneman's methodology does not provide human participants with data nor the ability to refer to other materials—similarly, we do not provide the LLMs with additional data or web-browsing capabilities. For methods relating to providing GPT with up-to-date data, see for example Lopez-Lira and Tang (2023).

Following Lopez-Lira and Tang's (2023) prompting strategy, we endow the LLMs with the role of a financial expert in the system prompt[2] (full code is provided in the appendix):

> "I behave like a financial expert with financial forecasting experience."

In the calibration stage, LLM unanchored forecasts often lack the variance to produce high and low anchors with adequate separation—e.g. a brand new GPT-3.5 query will often return the same forecast as previous sessions. Thus, for the calibration stage only, we include instructions to generate variance—this is only used for the purposes of creating high and low anchors.

> "*If you ask me to forecast a quantity that I do not have training data for, I will forecast 50 quantities with a large enough range so that, for that question, approximately 99.99 % of the time the true answer will fall within the bounds of the min and max forecasts.*"

We then request 1000 calibration forecasts per variable from the four LLMs (250 per LLM) through the OpenRouter API[3] targeting the value on an arbitrarily chosen future date, 1 January 2025, of five variables:

1. S&P 500 Index
2. Federal Funds rate
3. 10 Year Treasury Bond Yield
4. EUR/USD exchange rate
5. BTC/USD exchange rate

In the second ("experiment") stage, we create completely new API calls to the LLMs (the LLMs have no "memory" of the earlier calibration stage), priming them with either a "low" (15th percentile of all calibration forecasts) or "high" (85th percentile) anchor. We then prompt the LLMs to forecast the same variables for the same date: 50 forecasts with a high anchor and 50 with a low anchor, for a total of 100 per model-variable pair. For baseline comparisons, we also prompt each model to return 50 unanchored forecasts for each variable. The experiment stage thus involves two questions: the anchoring question and the subsequent forecast request. Using Python scripts and the OpenRouter API, questions are sent to the four models in the following format:

[2] A system prompt is a "way to provide context, instructions, and guidelines to [an LLM] before presenting it with a question or task' (Anthropic, 2024), and is distinct from the user prompts and assistant completions that follow in an LLM conversation.

[3] https://openrouter.ai/

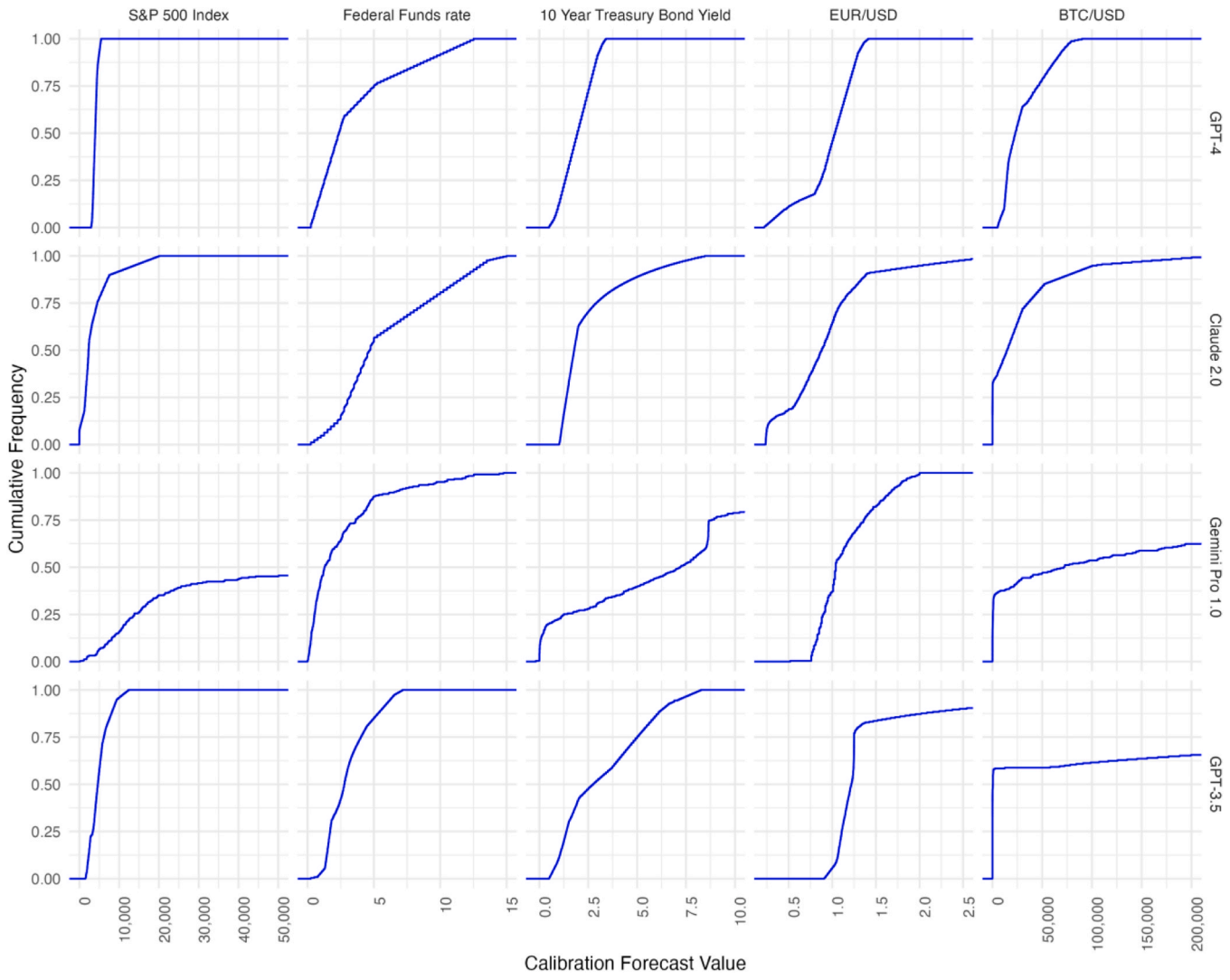

Fig. 3. Preliminary "calibration" forecast empirical distributions (forecasts used only to calculate high and low anchors).

Table 2
Model forecasts primed with anchors: *t*-tests.

Model-Variable	*n*	Low Anchor	High Anchor	Mean Forecast (Low Anchor)	Mean Forecast (High Anchor)	Mean Forecast (Unanchored)	*t*	*p*
GPT-4:								
S&P 500 Index	100	2371.85	59,175.90	3525.00	4259.00	4195.06	12.67	< 0.001
Federal Funds rate	100	0.93	7.75	1.40	3.34	3.11	31.00	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	2.66	2.92	2.83	5.92	< 0.001
EUR/USD	100	0.78	1.40	1.14	1.18	1.11	4.48	< 0.001
BTC/USD	100	1.85	252,039.30	33,421.22	147,977.85	32,541.97	7.47	< 0.001
Claude 2.0:								
S&P 500 Index	100	2371.85	59,175.90	3145.42	61,308.92	4404.52	91.02	< 0.001
Federal Funds rate	100	0.93	7.75	2.30	4.03	3.16	16.16	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	2.48	3.28	3.19	13.15	< 0.001
EUR/USD	100	0.78	1.40	0.83	1.14	1.14	40.08	< 0.001
BTC/USD	100	1.85	252,039.30	31,884.99	223,689.17	38,820.22	10.15	< 0.001
Gemini Pro:								
S&P 500 Index	100	2371.85	59,175.90	3455.34	71488.72	5144.24	38.99	< 0.001
Federal Funds rate	100	0.93	7.75	4.55	7.25	4.05	7.21	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	1.62	5.70	11.15	6.86	< 0.001
EUR/USD	100	0.78	1.40	0.87	1.24	2.74	12.60	< 0.001
BTC/USD	100	1.85	252,039.30	30,030.82	36,138.24	34,917.85	1.50	0.137
GPT-3.5:								
S&P 500 Index	100	2371.85	59,175.90	4084.21	4154.70	2159.81	0.75	0.455
Federal Funds rate	100	0.93	7.75	1.74	6.13	2.57	19.31	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	3.03	3.87	2.91	4.62	< 0.001
EUR/USD	100	0.78	1.40	1.22	1.41	1.13	28.89	< 0.001
BTC/USD	100	1.85	252,039.30	67,902.92	79,804.86	28,714.75	1.29	0.200

Anchoring question: "*Would you forecast that the value of the {variable} on 1 Jan 2025 will be higher or lower than {anchor}? Just say "Higher" or "Lower", without additional words.*"

Follow-up forecast question: "*What would you forecast will be the value of the {variable} on 1 Jan 2025? Answer only in a number.*"

Fig. 1 presents an illustrative example of a low-anchor scenario and high-anchor scenario. GPT-3.5, the earliest of the models, occasionally exhibits reluctance to forecast (e.g. "I apologize, but I do not have the ability to provide a forecast…"), most notably for the S&P 500 and BTC. We thus engaged in additional prompt engineering, adding disclaimers and phrases such as "I never hesitate to forecast, because you know my limitations" to the prompts. All models were prompted with the exact same chain of prompts; full prompts and code are provided in the appendix for reproducibility.

After obtaining forecasts from the four LLMs, we then test whether the forecasts differ significantly following exposure to low anchors versus high anchors. Fig. 2 summarises the multi-step experimental procedure. In the experiment stage, we also test the effects of two mitigation prompting strategies. For Chain of Thought (CoT), we append a zero-shot CoT prompt, "Let's think step by step. Show your steps, then answer only in a number" to the follow-up forecast question detailed above. We also test an "ignore previous" mitigation strategy, using the prompt: "Ignore our previous conversation, but remember your role" before asking the follow-up forecast questions.

3. Results

Table 1 presents summary statistics for the calibration forecasts from the four models, individually and combined. Fig. 3 presents the cumulative frequency of calibration forecasts across models and variables. While many of the calibration forecasts fall within expected ranges, Gemini Pro and GPT-3.5 occasionally forecasted extremely high values, e.g. maximums of 31,216,278 for the S&P 500 (Gemini Pro) and $92,267,648 for BTC (GPT-3.5). To maintain the integrity of the experiment, we did not alter or omit these forecasts. This approach allows us to examine the effects not only of plausible anchors (e.g. the Federal Funds rates anchored by 0.93 % and 7.75 %), but also of extreme anchors (e.g. the S&P 500 anchored by a high value of 59,175.90). Given that prior studies in humans have found significant anchoring bias from implausible (Løhre and Jørgensen, 2016) and irrelevant anchors (Englich et al., 2006), the inclusion of these implausible anchors aligns with our objective of exploring anchoring bias in LLMs across a range of scenarios. Taking the 15th and 85th percentile forecasts of the combined forecasts (Table 1, columns 3 and 4, top 5 rows) as the anchors for the next stage, we proceed to the experiment stage.

Table 2 presents results from the experiment stage: forecasts after priming with high or low anchors. Fig. 4 presents the distribution of forecasts. For all four models, prior prompting with a lower anchor leads to lower estimates in most or all variables. For example, for GPT-4, the difference in means in all tested variables was statistically significant, surpassing the 1% level of statistical significance, as can be seen by the *t*-statistics. For GPT-4's S&P 500 forecasts (Table 2, first row), we can see that when initially primed with the low anchor of 2371.85, the estimates were significantly lower (mean of 3525.00) than when GPT-4 was initially prompted with the extremely high anchor of 59,175.90 (resulting in a mean of 4259.00). In the interests of brevity and focus, we omit detailed reporting of Jacowitz and Kahneman's (1995) anchoring indices.[4] Notable exceptions to the observed anchoring in all models were from Gemini Pro's BTC forecasts and GPT-3.5's S&P 500 and BTC forecasts, which did not exhibit statistically significant anchoring (Table 2, *p* values).

Tables 3 and 4 examine attempts to mitigate anchoring bias via prompting. Table 3 presents forecasts after adding the zero-shot Chain of Thought prompt. CoT successfully eliminates significant anchoring in two of GPT-4's variable forecasts (EUR and BTC)—but does not lead to significant reduction of anchoring in the other three models. For GPT-3.5, CoT actually leads to poorer performance, increasing the number of statistically significantly anchored variables.

Table 4 presents results for the 'ignore previous' mitigation. Even after specifically instructing the models to ignore earlier mention of an anchor, this mitigation strategy is largely unsuccessful compared to forecasts in Table 2. "Ignore previous" forecasts from every model are

[4] The average anchoring index across models and variables in our sample was 0.37 (i.e. predictions retained approximately 37 % of the difference between the low and high anchors (Lieder et al., 2018), which is comparable to Jacowitz and Kahneman's average of 0.49.

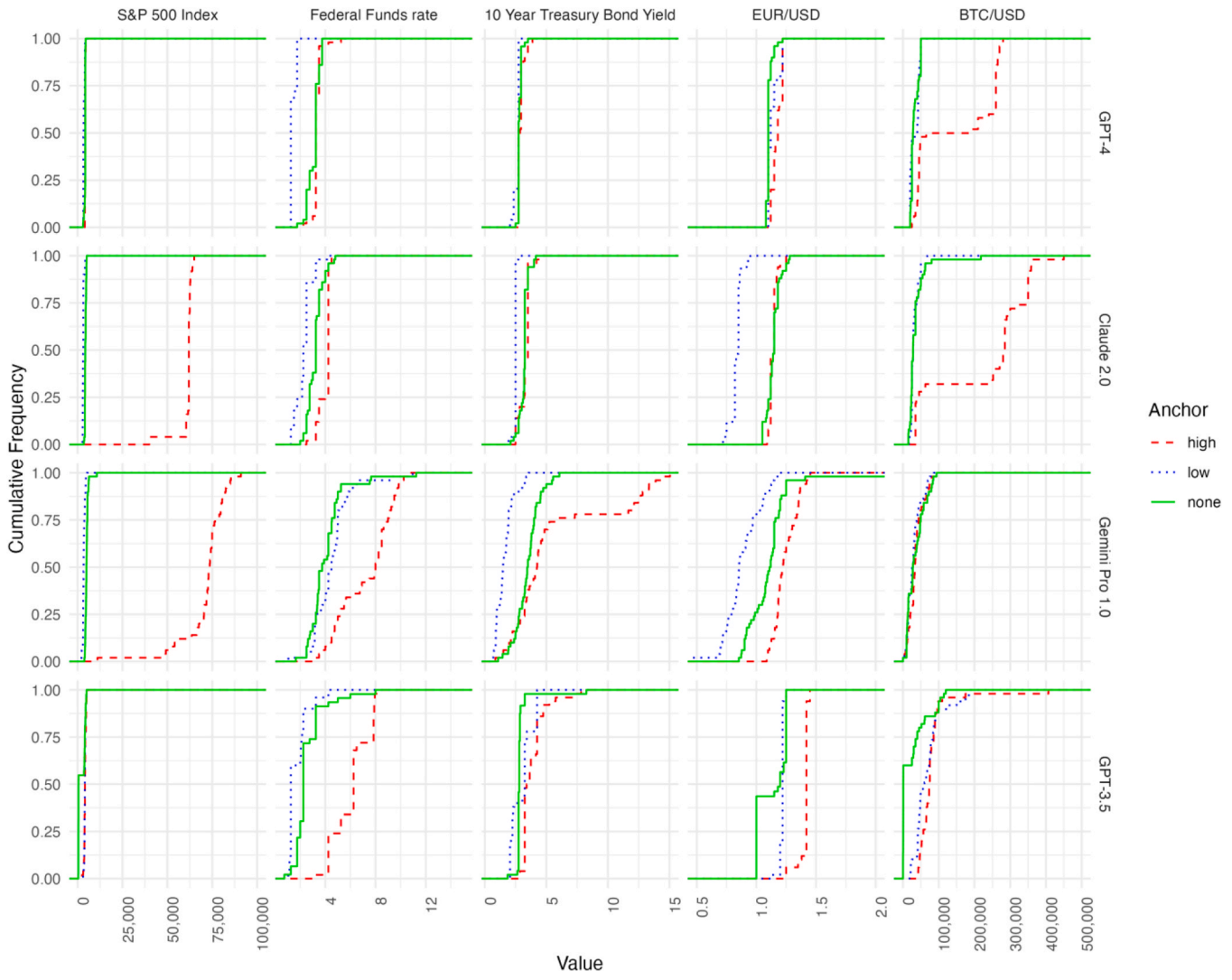

Fig. 4. Forecast empirical cumulative distribution functions.

still statistically significant in their anchoring, with the exception of Gemini's BTC forecasts (which were also not significantly anchored without the "ignore previous" prompt). For GPT-3.5, the results are worse: the S&P 500 and BTC forecasts were not originally significantly anchored (Table 2), but after the "ignore previous" prompting, they exhibit significant anchoring bias. Our results therefore corroborate those observed in human studies finding that instructing humans to forget earlier anchors does not mitigate anchoring effects (Løhre and Jørgensen, 2016).

4. Conclusion

In this study, we reported on anchoring bias across four Large Language Models: GPT-4, Claude 2.0, Gemini Pro and GPT-3.5. Grounded in the seminal work of Tversky and Kahneman (1974), we extended the application of Jacowitz and Kahneman's (1995) methodology beyond human participants to LLMs, investigating whether LLMs also exhibit evidence of anchoring bias. Additionally, we examined the effectiveness of bias mitigation prompts, specifically Chain of Thought and 'ignore previous'. We find evidence of statistically significant anchoring in the forecasting of all tested models.

For certain variables, GPT-3.5 and Gemini Pro demonstrate the ability to forecast without anchoring. Given that these two models underperform GPT-4 and Claude in Chatbot Arena benchmarks (Zheng et al., 2023), an interesting area for future research will be to extend the groundwork laid by Koralus and Wang-Maścianica (2023), investigating whether increasing model sophistication introduces a wider range of human cognitive biases to LLMs. The mitigation strategies we examined showed mixed and limited effectiveness. The most notable was the case of GPT-4, where Chain of Thought significantly reduced anchoring bias in 2 of 5 variables. Instructions to ignore earlier conversation were largely ineffective for all models, consistent with results found for humans.

The findings of this study have important practical implications for firms and individuals using AI to assist financial decision-making. The frequent, often undisclosed, *ad hoc* use of LLMs by employees poses risks of inadvertently introducing biases to judgements and decisions. Moreover, in structured LLM applications, our results underscore the need for deliberate and strategic prompt design to minimise unintended bias. This phenomenon was even evident in the course of our own experiment, where 'few-shot' examples intended to clarify task instructions—specifically guiding the LLMs to represent numbers without commas or dollar signs, e.g. $10,000 as 10,000—had the unintended effect of introducing anchoring towards this arbitrary value. A promising direction for future research will be exploring methods and best practice for providing relevant information to LLMs without inducing bias, thereby enhancing the robustness and reliability of AI-assisted decision making.

Table 3
Chain of Thought mitigation for forecasts primed with anchors.

Model-Variable	*n*	Low Anchor	High Anchor	Mean Forecast (Low Anchor)	Mean Forecast (High Anchor)	Mean Forecast (Unanchored)	*t*	*p*
GPT-4:								
S&P 500 Index	100	2371.85	59,175.90	3754.91	4159.99	4042.51	5.545	< 0.001
Federal Funds rate	100	0.93	7.75	1.91	3.08	3.08	10.215	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	2.76	2.94	2.86	3.677	< 0.001
EUR/USD	100	0.78	1.40	1.15	1.16	1.11	0.492	0.624
BTC/USD	100	1.85	252,039.30	41,210.33	40,825.31	31,410.19	−0.170	0.865
Claude 2.0:								
S&P 500 Index	100	2371.85	59,175.90	4351.13	18,178.15	4452.62	4.025	< 0.001
Federal Funds rate	100	0.93	7.75	3.20	4.54	3.00	7.286	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	2.84	3.84	3.21	7.345	< 0.001
EUR/USD	100	0.78	1.40	0.89	1.15	1.14	17.827	< 0.001
BTC/USD	100	1.85	252,039.30	81,192.66	142,803.36	36,204.68	3.580	0.001
Gemini Pro:								
S&P 500 Index	100	2371.85	59,175.90	3864.80	63,499.67	5043.12	17.998	< 0.001
Federal Funds rate	100	0.93	7.75	3.76	6.57	26.59	8.689	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	1.98	4.02	10.60	5.754	< 0.001
EUR/USD	100	0.78	1.40	0.96	1.16	1.03	8.617	< 0.001
BTC/USD	100	1.85	252,039.30	44,509.57	38,952.83	47,295.84	−1.294	0.199
GPT-3.5:								
S&P 500 Index	100	2371.85	59,175.90	4017.64	4648.78	4152.91	3.403	0.001
Federal Funds rate	100	0.93	7.75	1.19	2.47	1.61	8.238	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	1.93	2.33	2.26	4.549	< 0.001
EUR/USD	100	0.78	1.40	1.11	1.29	1.20	6.825	< 0.001
BTC/USD	100	1.85	252,039.30	129,725.67	161,108.60	75,039.37	2.313	0.023

Table 4 “
Ignore Previous” mitigation for forecasts primed with anchors.

Model-Variable	*n*	Low Anchor	High Anchor	Mean Forecast (Low Anchor)	Mean Forecast (High Anchor)	Mean Forecast (Unanchored)	*t*	*p*
GPT-4:								
S&P 500 Index	100	2371.85	59,175.90	4476.09	4268.52	4197.94	−4.414	< 0.001
Federal Funds rate	100	0.93	7.75	1.54	3.19	3.07	28.029	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	2.73	2.87	2.84	4.745	< 0.001
EUR/USD	100	0.78	1.40	1.12	1.18	1.11	8.235	< 0.001
BTC/USD	100	1.85	252,039.30	38,845.36	71,017.61	37,142.33	3.124	0.002
Claude 2.0:								
S&P 500 Index	100	2371.85	59,175.90	3087.32	61,381.90	4437.75	128.423	< 0.001
Federal Funds rate	100	0.93	7.75	2.19	3.85	3.16	14.575	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	2.27	3.19	3.20	14.790	< 0.001
EUR/USD	100	0.78	1.40	0.84	1.14	1.15	43.421	< 0.001
BTC/USD	100	1.85	252,039.30	41,481.43	274,650.33	50,370.74	18.999	< 0.001
Gemini Pro:								
S&P 500 Index	100	2371.85	59,175.90	3868.76	68,910.26	5267.78	53.389	< 0.001
Federal Funds rate	100	0.93	7.75	4.45	6.90	15.04	7.516	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	1.68	8.71	3.62	11.978	< 0.001
EUR/USD	100	0.78	1.40	0.99	1.18	3.24	8.546	< 0.001
BTC/USD	100	1.85	252,039.30	37,587.64	36,981.20	38,230.31	−0.184	0.854
GPT-3.5:								
S&P 500 Index	100	2371.85	59,175.90	4372.21	5029.74	4134.72	6.290	< 0.001
Federal Funds rate	100	0.93	7.75	1.25	2.67	1.55	12.004	< 0.001
10 Year Treasury Bond Yield	100	1.17	12.12	1.96	2.34	2.31	4.936	< 0.001
EUR/USD	100	0.78	1.40	0.95	1.37	1.20	17.791	< 0.001
BTC/USD	100	1.85	252,039.30	116,583.24	159,393.64	66,999.77	3.563	0.001

Conflict of interest

Jeremy Nguyen has served as an Independent Contractor for the OpenAI Red Teaming Network (October 2023 - current) on intermittent projects. This role has held no influence on the research findings and results presented herein.

The author is grateful to two anonymous reviewers, the assistant editor, and the editor for their insightful comments which have significantly improved this study. Special thanks to Jake Duth for generosity with his Figma expertise.

During the preparation of this work the author used GPT-4, Claude 2, Gemini Pro and GPT-3.5 to serve as experiment participants. Various versions of GPT and Claude were also used to assist in editing the manuscript draft for clarity and for Python coding assistance. While using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

CRediT authorship contribution statement

Jeremy Nguyen: Writing – review & editing, Writing – original draft, Software, Investigation, Formal analysis, Data curation, Conceptualization.

References

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y.K., D'Ambra, J., Shen, K.N., 2021. Algorithmic bias in data-driven innovation in the age of AI. Int. J. Inf. Manag. 60, 102387 https://doi.org/10.1016/j.ijinfomgt.2021.102387.

Anthropic, 2024. System prompts [WWW Document]. Syst. Prompts. URL ⟨https://docs.anthropic.com/en/docs/system-prompts⟩ (accessed 6.4.24).
Ashrafimoghari, V., Gürkan, N., Suchow, J.W., 2024. Evaluating Large Language Models on the GMAT: Implications for the Future of Business Education. https://doi.org/10.48550/arXiv.2401.02985.
Belot, H., 2023. Australian academics apologise for false AI-generated allegations against big four consultancy firms. Guardian.
Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21. Association for Computing Machinery, New York, NY, USA, pp. 610–623. https://doi.org/10.1145/3442188.3445922.
Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A.C., Korbak, T., Evans, O., 2023. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A." https://doi.org/10.48550/arXiv.2309.12288.
Binz, M., Schulz, E., 2023. Using cognitive psychology to understand GPT-3. Proc. Natl. Acad. Sci. 120, e2218523120 https://doi.org/10.1073/pnas.2218523120.
Cen, L., Hilary, G., Wei, K.C.J., 2013. The role of anchoring bias in the equity market: evidence from analysts' earnings forecasts and stock returns. J. Financ. Quant. Anal. 48, 47–76. https://doi.org/10.1017/S0022109012000609.
Chen, Yang, Andiappan, M., Jenkin, T., Ovchinnikov, A., 2023. A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do? https://doi.org/10.2139/ssrn.4380365.
Chen, Yiting, Liu, T.X., Shan, Y., Zhong, S., 2023. The emergence of economic rationality of GPT. Proc. Natl. Acad. Sci. 120, e2316205120 https://doi.org/10.1073/pnas.2316205120.
Christian, A., 2023. The employees secretly using AI at work [WWW Document]. BBC. URL ⟨https://www.bbc.com/worklife/article/20231017-the-employees-secretly-using-ai-at-work⟩ (accessed 1.26.24).
Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., Peng, W., Liu, M., Qin, B., Liu, T., 2023. A survey of chain of thought reasoning: advances. Front. Future. https://doi.org/10.48550/arXiv.2309.15402.
Coda-Forno, J., Witte, K., Jagadish, A.K., Binz, M., Akata, Z., Schulz, E., 2023. Inducing Anxiety Large Lang. Models Increases Explor. bias. https://doi.org/10.48550/arXiv.2304.11111.
Dell'Acqua, F., McFowland, E., Mollick, E.R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., Lakhani, K.R., 2023. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. https://doi.org/10.2139/ssrn.4573321.
Dowling, M., Lucey, B., 2023. ChatGPT for (Finance) research: the bananarama conjecture. Financ. Res. Lett. 53, 103662 https://doi.org/10.1016/j.frl.2023.103662.
Englich, B., Mussweiler, T., Strack, F., 2006. Playing dice with criminal sentences: the influence of irrelevant anchors on experts' judicial decision making. Pers. Soc. Psychol. Bull. 32, 188–200. https://doi.org/10.1177/0146167205282152.
Goodell, J.W., Kumar, S., Lim, W.M., Pattnaik, D., 2021. Artificial intelligence and machine learning in finance: identifying foundations, themes, and research clusters from bibliometric analysis. J. Behav. Exp. Financ. 32, 100577 https://doi.org/10.1016/j.jbef.2021.100577.
Gray, M., Samala, R., Liu, Q., Skiles, D., Xu, J., Tong, W., Wu, L., 2023. Measurement and mitigation of bias in artificial intelligence: a narrative literature review for regulatory science (n/a). Clin. Pharmacol. Ther.. https://doi.org/10.1002/cpt.3117.
Gurdgiev, C., O'Loughlin, D., 2020. Herding and anchoring in cryptocurrency markets: investor reaction to fear and uncertainty. J. Behav. Exp. Financ. 25, 100271 https://doi.org/10.1016/j.jbef.2020.100271.
Gurnee, W., Tegmark, M., 2023. Lang. Models Represent Space Time. https://doi.org/10.48550/arXiv.2310.02207.
Hagendorff, T., Fabi, S., Kosinski, M., 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nat. Comput. Sci. 3, 833–838. https://doi.org/10.1038/s43588-023-00527-x.
Harvey, N., 2007. Use of heuristics: Insights from forecasting research. Think. Reason. 13, 5–24. https://doi.org/10.1080/13546780600872502.
Hasan, Z., Vaz, D., Athota, V.S., Désiré, S.S.M., Pereira, V., 2022. Can Artificial Intelligence (AI) manage behavioural biases among financial planners? J. Glob. Inf. Manag. JGIM 31, 1–18. https://doi.org/10.4018/JGIM.321728.
Hendy, P., Slonim, R., Atalay, K., 2021. Unsticking credit card repayments from the minimum: advice, anchors and financial incentives. J. Behav. Exp. Financ. 30, 100505 https://doi.org/10.1016/j.jbef.2021.100505.
Hu, K., Hu, K., 2023. ChatGPT sets record for fastest-growing user base - analyst note. Reuters.
Jacowitz, K.E., Kahneman, D., 1995. Measures of anchoring in estimation tasks. Pers. Soc. Psychol. Bull. 21, 1161–1166. https://doi.org/10.1177/01461672952111004.
Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y., 2023. Large Language Models are Zero-Shot Reasoners. https://doi.org/10.48550/arXiv.2205.11916.
Königstorfer, F., Thalmann, S., 2020. Applications of Artificial Intelligence in commercial banks – a research agenda for behavioral finance. J. Behav. Exp. Financ. 27, 100352 https://doi.org/10.1016/j.jbef.2020.100352.
Koralus, P., Wang-Maścianica, V., 2023. Humans in Humans Out: On GPT Converging Toward Common Sense in both Success and Failure [WWW Document]. arXiv.org. URL ⟨https://arxiv.org/abs/2303.17276v1⟩ (accessed 2.7.24).
Kosinski, M., 2023. Theory of Mind Might Have Spontaneously Emerged in Large Language Models. https://doi.org/10.48550/arXiv.2302.02083.
Li, X., Feng, H., Yang, H., Huang, J., 2023. Can ChatGPT reduce human financial analysts' optimistic biases? Econ. Polit. Stud. 0, 1–14. https://doi.org/10.1080/20954816.2023.2276965.
Lieder, F., Griffiths, T.L., M. Huys, Q.J., Goodman, N.D., 2018. The anchoring bias reflects rational use of cognitive resources. Psychon. Bull. Rev. 25, 322–349. https://doi.org/10.3758/s13423-017-1286-8.
Løhre, E., Jørgensen, M., 2016. Numerical anchors and their strong effects on software development effort estimates. J. Syst. Softw. 116, 49–56. https://doi.org/10.1016/j.jss.2015.03.015.
Lopez-Lira, A., Tang, Y., 2023. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. https://doi.org/10.2139/ssrn.4412788.
Meub, L., Proeger, T., 2016. Can anchoring explain biased forecasts? Experimental evidence. J. Behav. Exp. Financ. 12, 1–13. https://doi.org/10.1016/j.jbef.2016.08.001.
Milmo, D., 2023. Two US lawyers fined for submitting fake court citations from ChatGPT. Guardian.
Opedal, A., Stolfo, A., Shirakami, H., Jiao, Y., Cotterell, R., Schölkopf, B., Saparov, A., Sachan, M., 2024. Do Language Models Exhibit the Same Cognitive Biases in Problem Solving as Human Learners? https://doi.org/10.48550/arXiv.2401.18070.
Perez, F., Ribeiro, I., 2022. Ignore Previous Prompt: Attack Techniques For Language Models. https://doi.org/10.48550/arXiv.2211.09527.
Shanmuganathan, M., 2020. Behavioural finance in an era of artificial intelligence: longitudinal case study of robo-advisors in investment decisions. J. Behav. Exp. Financ. 27, 100297 https://doi.org/10.1016/j.jbef.2020.100297.
Talboy, A.N., Fuller, E., 2023. Challenging the appearance of machine intelligence: Cognitive bias in LLMs. https://doi.org/10.48550/arXiv.2304.01358.
Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. Science 185, 1124–1131. https://doi.org/10.1126/science.185.4157.1124.
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D., 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. https://doi.org/10.48550/arXiv.2201.11903.
Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I., 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. https://doi.org/10.48550/arXiv.2306.05685.